Critical mutation rates in small populations

Elizabeth Jane Aston

Submitted for the degree of

Doctor of Philosophy

June 2014

Keele University

Acknowledgements

There are a number of people who have helped me greatly and whose support I could not have done without. I am very grateful to my supervisor Alastair Channon for his guidance over the course of my PhD. His insight, ideas and technical support have helped me to develop my own ideas and research techniques. My confidence as a researcher has increased greatly thanks to his support. I am also grateful for the additional support of my second supervisor Charles Day for his stimulating discussion and for providing some very useful alternative opinions and suggestions.

Many thanks go to Roman Belavkin, Chris Knight, Rok Krasovec and John Aston for providing an interesting experience working as part of a multidisciplinary project team. Their additional input has sparked ideas that would not have otherwise arisen, and provided expertise in each of their relevant fields. Along with my supervisors they have provided discussion and guidance relating to results and development of further work. Chris Knight provided useful assistance with using R as a curve-fitting tool. I am grateful to the EPSRC for the funding, and for the opportunity to work with and have support from such a wide range of people. Thanks also to my examiners Dov Stekel and Gordon Rugg for providing constructive feedback on my thesis and allowing me the opportunity to discuss my work as a peer.

I would not have been able to dedicate myself to this work without the emotional support of my family and friends. In particular, I would like to thank my parents for

putting up with me using them as a sounding board for my ideas, for their encouragement, and for keeping me going during the difficult times. I would also like to thank Dave for listening to me and for being patient with me. His willingness to listen to my ideas and take an interest helped me to keep up my motivation; his patience and understanding have been invaluable.

Abstract

Mutation introduces change at the sequence level. There is a critical mutation rate above which changes occur too frequently for natural selection to maintain the population's genetic makeup. This thesis examines the relationship between this critical mutation rate and the number of individuals in the adapting population. It presents an algorithmic method capable of providing widely applicable results in haploid and diploid populations, and verifies this method against analytical models for the error threshold.

Use of the method led to the discovery of an exponential relationship between the critical mutation rate and population size, particularly strong in small populations with 100 individuals or less, contradicting the existing idea that critical mutation rate and population size are independent [1]. The critical mutation rate (and error threshold) were found to be lower in diploids due to differences in recombination. Analysis of the survival-of-the-fittest to survival-of-the-flattest transition enabled improvement of existing definitions of the critical mutation rate.

Development of a faster algorithm capable of running experiments with parameter values within the range found in nature began the process of bridging the gap between artificial and biological evolution. A link was established between the exponential model and natural mutation rates. Increasing the gene length by a factor of 10 was found to decrease both the critical mutation rate and error threshold by an order of

magnitude. Natural mutation rates lie below these values, although further work is required to establish any trend. A potential link has been established between the critical mutation rate, error threshold, and optimal mutation rate control theory.

Future work may develop the algorithmic method to include more complex features of biological populations, and go on to determine the effect the exponential model can have on population extinction, recovery, and conservation.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Small populations frequently exist in nature. Those nearing extinction may contain no more than a few individuals. For example, the Chatham Island Black Robin was recorded as existing in two populations consisting of 190 and 34 mature individuals as of spring 2011 [4]. As of 2008, the total worldwide known cheetah population consisted of approximately 7500 adults. However, only four of the 15 known populations of cheetahs in Eastern Africa were estimated to consist of greater than 200 individuals [5]. Based on sampling of adult individuals, Campbell's Alligator Lizard has been reported as having a total estimated population size of 500 individuals as of 2010 [6]. Environmental change is rapid, therefore populations need to evolve at a sufficient rate to prevent further population decline and enable evolutionary rescue [7]. Further population decline can lead to loss of fit genetic material that may be difficult to recover in very small populations due to mutational meltdown [8]. Meltdown occurs when a deleterious mutation becomes fixed in a population leading to reduced fitness and therefore reduction in population size. Mutations become fixed more rapidly the fewer individuals there are in the population; each time fixation of a deleterious mutation leads to reduction in population size it becomes easier for further deleterious mutations to become fixed

leading to a potential downward spiral towards extinction. Understanding the effect of population size on the critical parameters of evolution (mutation, recombination, selection, and genetic drift) is essential in making accurate predictions regarding the likely fate of such a population.

Evolution occurs through the process of mutation, recombination, selection and genetic drift in accordance to the fitness landscape. The concept of a fitness landscape was introduced by Wright [9] and later combined with the notion of sequence space by Eigen and Schuster [10]. Each sequence in sequence space has a fitness value, which represents its relative replication capacity [11]. The fittest sequences in the landscape are the 'peaks', while the lower fitness sequences occupy the 'valleys'. Mutation introduces variation, while selection acts to increase the frequency of fitter sequences. The balance between these two forces is referred to as the mutation-selection balance [12, 13]. When there is mutation-selection balance, the population will tend to cluster around the fitness peaks and form a quasispecies [10, 13, 14]. A quasispecies is a well-defined distribution of mutants generated by a mutation-selection process [15].

The degree to which a genetic perturbation affects fitness is dependent on the robustness of an individual's genotype. Robustness is defined as the average effect of a specific type of perturbation (such as a new mutation) on the fitness of a specific genotype [16]. The greater the robustness, the smaller the change in fitness of a genotype after mutation. In mammals, the majority of mutations have no effect on the phenotype of the individual, i.e., most mutations are neutral [17]. However, most non neutral mutations are detrimental to fitness [18], therefore robustness can limit the damage each time a mutation occurs. Conversely, sensitivity to mutation is an advantage in the rare case of beneficial mutation; there is persistant pressure to evolve sequences that are both fit and robust [19, 20, 21]. Sometimes individuals with greater

robustness to mutation are favoured over individuals with greater fitness, a concept known as 'survival-of-the-flattest'.

1.1 Population Size and the Evolutionary Process

The efficiency of natural selection is positively dependent on the size of a population [22, 23], the reason being that small populations are more influenced by genetic drift and therefore have a greater chance of deleterious mutations becoming fixed [24, 23]. Beneficial mutations are also less likely to occur when the population is small [23]. According to Wright's Shifting Balance Theory, in which movement to a higher peak occurs in three phases (genetic drift, natural selection, and outcompeting or interbreeding with other subpopulations to move the global population) [25, 9], evolution occurs more quickly when a population divides into smaller subpopulations as each small subpopulation will experience greater genetic drift than the population as a whole.

In a landscape with a single fitness peak, a quasispecies is able to maintain its position surrounding the top of the peak so long as the mutation rate does not exceed a particular rate known as the error threshold. Above this threshold, there is an error catastrophe and the population delocalizes across sequence space [26, 13, 14, 27, 11, 28, 29]. Error thresholds also exist in multiple-peak landscapes [30]. The concept of the error threshold was introduced by Eigen [31] and later described by Nowak and Schuster [32]. It is dependent on the existence of the mutation-selection balance, and is the maximal mutation rate that allows a population to stay clustered around the fitness peak. In addition to the error threshold, in landscapes where there is more than one peak, there may also be one or more critical mutation rates at which the population loses its ability to remain on fitter peaks, but retains its ability to remain on flatter peaks of lower fitness [33, 26, 1, 34]. Above such critical mutation rates,

individuals with greater robustness to mutation are able to survive while fitter, less robust individuals may not. This represents a phase transition from survival-of-the-fittest to survival-of-the-flattest [33, 13, 1, 34, 21, 35, 29].

Krakauer and Plotkin [36] suggest that both in theory and in individual-based stochastic simulations, robustness increases the mean fitness in small populations as it masks mutations that arise due to mutational drift. However, large populations are less affected by drift, and so are more able to occupy high-fitness peaks in sharp landscapes. Both Wilke [37] and Comas [1] found "that population size played only a minor role in determining the position of the critical mutation rate" [34], within the context of their experiments. Comas [1] used population sizes as low as 250 and concluded "that the critical mutation rate was independent of population size" despite the fact that there did appear to be some correlation for certain cases. They did not consider smaller populations, such as those that may exist for species nearing extinction or living in localized groups. Both Nowak and Schuster [32] and Wiehe [38] considered the effect of random genetic drift in finite populations (in haploids and diploids respectively), and observed that there is a shift of error thresholds to lower values which is more pronounced the smaller the population. Error thresholds were also shown to increase for increasing population size using a genetic algorithm [39] with both single-peak and correlated landscapes [40]. Based on these results for error thresholds, consideration of the critical mutation rate when the population size is small may provide new insights into the effect of mutation on population decline.

1.2 Aims

The purpose of this work is to use genetic algorithms to examine the significance of survival-of-the-flattest in small populations. Specifically, it aims to:

1. Examine the relationship between population size and the critical mutation rate at which individuals with greater robustness to mutation are favoured over individuals with greater fitness (survival-of-the-flattest).

2. Test if the relationship between population size and the critical mutation rate holds for a diploid population modelled on the biological process of meiosis.

3. Examine the effect of haploidy and diploidy on the critical mutation rate.

4. Use parameter values from nature to verify that the simulation models used and the results obtained have relevance to biological systems.

5. Relate the critical mutation rate and the error threshold to the notion of optimal mutation rate control.

These aims have been achieved by:

1. Designing a simulation model to be implemented as a genetic algorithm (GA). This model must be capable of producing widely applicable results that do not rely precisely on the underlying fitness landscape used in the GA.

2. Developing this simulation model, moving closer to a biologically-inspired algorithm, and with the inclusion of diploidy.

3. Comparing both the algorithm and results obtained from the haploid and diploid methods.

4. Carrying out a search of the biological literature to identify ranges of equivalent GA parameter values observed in nature.

5. Comparing the critical mutation rate and error threshold results produced by the GA with mutation rates observed in nature and with optimal mutation rate control theory.

1.3 Organisation

This thesis is organised as follows:

Chapter 2 introduces the basic concepts of evolution and adaptation. It covers key ideas and equations in the broader areas of evolutionary dynamics and population genetics.

Chapter 3 goes on to focus on mutation and its role in adaptation. Specifically, it covers the different types of mutation, the nature of adaptation, optimal mutation rates and error thresholds, and the concept of mutational robustness and survival-of-the-flattest. Finally, factors associated with small populations and the risk of extinction are discussed.

Chapter 4 introduces the main aims and specific research questions. It then describes the simulation models used to address these questions. This chapter provides an introduction to the aims and methodology behind this work, which is elaborated on in subsequent chapters in conjunction with the outcomes.

Chapter 5 discusses the critical mutation rate with a review of key findings and studies, leading to the formation of the hypothesis that critical mutation rate has a dependence on population size in haploid populations. The relationship between the critical mutation rate, error threshold, and population size is studied using a two-peak landscape, and results produced using the haploid simulation model are presented. Verification of the error threshold results, and therefore the model itself, is provided by demonstrating consistency between the results and existing analytical models of error

threshold.

Chapter 6 uses the findings presented in chapter 5, in conjunction with previous studies of error thresholds in diploid populations, to form the hypothesis that critical mutation rate has a dependence on population size in diploid populations. Results are presented that examine the relationship between the critical mutation rate, error threshold, and population size using a two-peak landscape and the diploid version of the simulation model. The change in recombination systems when moving from haploidy to diploidy is also examined.

Chapter 7 begins to bridge the gap between artificial and biological evolution. The hypothesis is formed that increasing the sequence length will lower both the critical mutation rate and error threshold in line with the exponential model, and that neither the critical mutation rate nor the error threshold will go below the typical mutation rates found in nature. Results are presented that confirm the hypothesis by examining the relationship between critical mutation rate, error threshold, and population size using a two-peak landscape and the diploid version of the simulation model using parameter values found in nature.

Chapter 8 examines work relating to optimal mutation rate control, its relation to the error threshold, and its potential relation to the critical mutation rate.

Chapter 9 presents a summary of the achievements and key insights of this work, its limitations, and a discussion of potential future work.

This is followed by the appendices which contain copies of the code used to run the experiments, and a glossary of key terms.

1.4 Published Work and Attributions

Parts of this thesis have been published elsewhere, specifically:

- Most of chapter 5 has been published in [41].
- Most of chapters 5 and 6 has been published in [42].
- Work on optimal mutation rates associated with chapter 8 has either been published in [43] or is under review.

Work done explicitly by Elizabeth Aston:

- Identification of the critical mutation rate as a suitable candidate for study.
- Development of associated hypotheses.
- Implementation of haploid genetic algorithm based on established methods.
- Design and implementation of diploid genetic algorithm.
- Analysis of results and design of resulting further work.
- Discussion of optimal mutation rate control work (for which Elizabeth did not carry out the work, but was involved in some of the background reading and discussions and is therefore a contributing author) in terms of the results for the critical mutation rate.

Contributions by associates:

- Assistance in technical aspects of genetic algorithms.
- Use of R as a curve-fitting tool.
- Discussion and guidance relating to results and development of further work.

Chapter 2

Dynamics and the State of Evolving Systems

Populations are continually evolving and becoming better adapted to their current environment. This occurs by the action of processes that increase variation and select to maintain changes that positively affect the chance of survival. The greater an individual's chance of surviving, the greater the chance they will pass their genes on to the next generation, and so the greater their fitness. Mutation of the genetic sequences increases the amount of variation in the population, while selection favours those individuals that have high fitness. The theory of evolution by means of natural selection was put forward by Darwin and Wallace in 1858 [44], and further explained by Mendel's theory of inheritance in 1865 (although Mendel's work was not recognised until 1900) [45]. Previous to this, it was widely accepted that offspring were a fusion of their parents; it was believed that inherited characteristics were simply a blend of those found in the parents (known as blending theory). One of the major problems with this was the notion that, due to each offspring inheriting characteristics that must be somewhere in between those of its parents, the range of possible characteristics would decrease over

generations; blending inheritance would ultimately lead to uniformity. Darwin noted that, for there to be continual variation, the mechanisms by which new variation arise must always be working. He suggested that sexual reproduction would neutralise this new variation, unless selection acted to cause an accumulation of variation in a certain direction, and consequently a permanent evolutionary change [45]. Darwin's doubts about the once widely accepted concept of blending inheritance lead him to form the theory of natural selection, one of the most fundamental concepts in evolution [44, 46]. Another idea was the theory that an organism will pass on to its offspring any traits or characteristics acquired during its lifetime. This is known as Lamarckism (after Jean Baptiste Lamarck, who proposed the theory). August Weismann disputed this in his 1904 work entitled 'The Evolution Theory' [47]. He determined that there was no empirical evidence for Lamarckism, and that it was inconceivable that observable traits could be passed on directly through the germ line. However, recent studies have shown that epigenetic traits can be passed on to future generations [48, 49, 50]. Epigenetics refers to changes in gene activity that occur due to mechanisms that do not change the genetic code, for example, methylation of DNA or RNA interference [51, 50]. This bears resemblance to the Lamarckist view that the environment plays a role in influencing the genotype to determine the phenotype, and forms the basis for a form of "soft-lamarckism" [50].

Although Lamarck's ideas regarding the theory of evolution were disputed, he is credited by Darwin as being the first individual whose conclusions on the subject excited much attention [46, 52]. Since then, much work has been done regarding evolution at the level of the genetic sequence, the level of the population, and on the mechanisms that cause variation. This chapter presents a review of the literature that introduces the fundamental principles of evolution. It describes the key findings,

covering areas including evolutionary dynamics, population genetics, and the concepts of mutation and adaptation.

2.1 Evolutionary Dynamics

DNA sequences code for amino acids, which are then joined together to form proteins. John Maynard Smith came up with the notion that all proteins of a given length could be positioned so that immediate neighbours would have only one amino acid difference. This is referred to as sequence space [14]. The number of points in the sequence space for proteins is 20^L, where L is the length of the protein sequence, and 20 is the number of different standard amino acids from which a protein can be made. The number of dimensions in the sequence space is equal to L. As sequences can be very long, sequence spaces will often have many dimensions. These dimensions correspond to evolutionary trajectories; the longer the sequences, the larger the number of possible directions that can be followed in sequence space. The number of protons in the observable universe is estimated to be around 10^{80}, and this number will on many occasions be exceeded by the number of points in sequence space; the number of possible proteins by far exceeds the number of protons in the universe, even for relatively short sequences. There is therefore a limit on the proportion of possible protein sequences that can be explored during evolution. The same concept also applies to DNA and RNA, the only difference being that nucleic acids have an alphabet size of 4, meaning the number of points in the sequence space is 4^L. Richard Hamming came up with the concept of Hamming distance to calculate the distance between sequences in sequence space, as opposed to Euclidean distance. Hamming distance measures the minimum number of substitutions it takes to get from one sequence to another [14].

Each sequence in sequence space has a fitness value which represents its relative

replication capacity [11]. The concept of a fitness landscape was invented by Sewall Wright, and later combined with the notion of sequence space by Manfred Eigen and Peter Schuster [14]. Fitness landscapes are sometimes considered to resemble mountain ranges, with the fittest sequences at the peaks. However, this concept translates poorly to high dimensional sequence spaces with a low alphabet size, for example, nucleic acids which have an alphabet size of four (in that they are sequences consisting of four possible units, A, C, G and T, the four bases of DNA). Exploration of sequence space is done through evolution by mutation and selection in accordance to the fitness landscape; selection increases the frequency of the fittest individuals, while mutation introduces variation which is often at a cost of individual fitness. The balance between these two forces is referred to as the mutation-selection balance [12, 13]. A population in mutation-selection balance will tend to cluster around the fitness peaks and form what is known as a quasispecies [10, 13, 14]. This is based on the idea that a species in chemistry is a group of identical molecules (and a quasispecies is therefore a group of molecules that are not identical, but are related). The quasispecies equation describes how a population moves through sequence space:

$$\dot{x_i} = \sum_{j=1}^{m} x_j f_j q_{ji} - \phi x_i \tag{2.1}$$

Here, x_i is the frequency of sequence number i, where $i \in \{1, \ldots, \alpha^L\}$, α is the alphabet size, L is the length of sequences, $\sum x_i = 1$, f_j is fitness (selection), $\phi = \sum x_i f_i$ is the average fitness, and q_{ji} is a transition probability (the probability that the replication of sequence j leads to the creation of sequence i according to the mutation rate q_{ji}). The rate of change is denoted $\dot{x}$, and there are $m = \alpha^L$ sequences. Fitness f_i represents the fitness of sequence i; sequence i is reproduced at a rate equal to f,

to produce genome j. The probability that replication of sequence j will lead to the production of sequence i is given by:

$$q_{ji} = u^{h_{ji}}(1-u)^{L-h_{ji}} \tag{2.2}$$

It should be noted that Equation 2.2 only applies when $\alpha = 2$. When $\alpha > 2$, mutation could be to a different letter to that in either sequence i or j. Point mutations occur when one base is replaced by another during replication, with u the probability of mutation occurring at a particular position in the sequence. This makes 1–u the probability of correct replication. The number of mutations it takes to get from sequence i to sequence j is equal to the Hamming distance, h_{ij}. There is an assumption that mutation at one position will not affect the probability of mutation at the next position [14].

2.1.1 Changing Traits: Evolutionary Game Theory

Evolutionary dynamics at the genotype (sequence) level can be described at the phenotype (trait) level using evolutionary game theory. This is used to represent natural selection of evolutionary strategies. In the context of evolutionary game theory, fitness of an individual is said to be dependent on the frequency of other strategies in the population. A population of individuals can be considered to be the players in a game. Each individual has a strategy, and there are random interactions between the individuals. The resulting payoffs represent fitness, and those individuals that are successful when playing the game are the ones that will have the most reproductive success. The best strategies will be reproduced, whereas the poorest will be beaten each time there is competition; this represents natural selection. Adaptive dynamics

describes the way in which continuous traits and strategies alter during mutation and frequency-dependent selection; it is often used when looking at strategies that, when used by the whole population, cannot be beaten by any other strategy that starts to emerge (evolutionary stable strategies) [14].

Let x_i represent the frequency of strategy i in the population. Vector X represents the distribution of the population, i.e., $X = (x_1, \ldots, x_n)$. The fitness of strategy i is a function of the distribution of the population, and is given as $f_i(X)$. The average fitness of the population is ϕ. The dynamics of evolutionary games between different phenotypes can be described using the replicator equation [53]:

$$\dot{x}_i = x_i[f_i(X) - \phi] \tag{2.3}$$

Fitness may be dependent on the frequency of a phenotype. As an example, consider phenotypes A and B, where individuals with phenotype A can move but B cannot. The ability to move may give A an advantage over B. However, this advantage may be lost when the population is highly dense, as A will become obstructed from moving around. The cost of the ability to move would normally be outweighed by the benefits of moving; if A cannot move, the cost will no longer be outweighed and B will have the advantage. In this case, selection is frequency-dependent while fitness is variable. This concept can be formalised within the replicator equation:

$$\dot{x_A} = x_A[f_A(X) - \phi]$$

$$\dot{x_B} = x_B[f_B(X) - \phi]$$

where $f_A(X)$ and $f_B(X)$ represent the fitness of A and B, and x_A and x_B represent the frequency of A and B respectively. The average fitness is

$$\phi = x_A f_A(X) + x_B f_B(X)$$

As the sum of the frequency of A and B will always be 1, a new variable x can be added, where $x_A = x$ and $x_B = 1 - x$. Using this new variable, it can be said that the fitness of A and B is $f_A(x)$ and $f_B(x)$ respectively [14].

The Lotka-Volterra equation is used to describe the interaction between n separate species in a population, where y_i represents the frequency of species i, and f_i represents its fitness [53]. It should be noted that in the context of this equation, fitness is defined as the reproductive rate of a species. The Lotka-Volterra equation is given as:

$$\dot{y_i} = y_i f_i(Y) \tag{2.4}$$

Neither the replicator equation nor the Lotka-Volterra equation include mutation, while the quasispecies equation does not include frequency-dependent selection. Selection and mutation are key to adaptation, therefore they should ideally both be included . This can be done by combining the quasispecies equation and the Lotka-Volterra equation to produce the replicator-mutator equation:

$$\dot{x_i} = \sum_{j=0}^{m} x_j f_j(X) q_{ji} - \phi x_i \tag{2.5}$$

The quasispecies and Lotka-Volterra equations can be used in special circumstances (i.e., where there is to be no frequency-dependent selection or mutation respectively), but the replicator-mutator equation is more suitable in general [53].

The Price equation describes any form of selection, and is given as

$$\dot{E}(p) = Cov(f, p) + E(\dot{p}) \tag{2.6}$$

$\dot{E}(p)$ is the change in the expected value of trait p. p_i is the numerical value of an (arbitrary) trait of an individual i from the population. The average value of this trait across the population is given as $\bar{p}$.

$$\bar{p} \equiv E(p) = \sum\nolimits_i p_i x_i$$

$Cov(f, p)$is the covariance of trait p with fitness f.

$$Cov(f, p) = \sum\nolimits_i x_i f_i p_i - \phi\bar{p}$$

If the trait p_i does not change over time, the Price equation can be simplified to give

$$\dot{E}(p) = Cov(f, p)$$

2.1.1.1 Equivalence of Equations

The replicator equation can be shown to be equivalent to the Price equation, and the replicator-mutator equation can be shown to be equivalent to an expanded version of the Price equation [53]. The expanded Price equation has an additional term:

$$\dot{E}(p) = Cov(f, p) + E(\dot{p}) + E(f\Delta_m p) \tag{2.7}$$

where $E(f\Delta_m p)$ describes mutation among types. $E(p) = \sum_i p_i x_i$, therefore

$$E(f\Delta_m p) = \sum\nolimits_i x_i f_i \Delta_m p_i$$

$\Delta_m p_i = \sum_j q_{ij}(p_j - p_i)$, and this gives the expected change in the value of trait p when there is a mutation from type i. $\dot{E}(p) = \sum_i p_i \dot{x_i} + \sum_i x_i \dot{p_i}$ (where $\dot{E}(p)$ is the change in expected value of trait p, $\dot{x_i}$ is the change in frequency of individuals with genome i, and $\dot{p_i}$ is the change in value of trait p of individual i).

The replicator-mutator equation states that $\dot{x_i} = \sum_{j=0}^{m} x_j f_j(X) q_{ji} - \phi x_i$. Therefore

$$\dot{E}(p) = \sum\nolimits_i p_i (\sum\nolimits_j x_j f_j q_{ji} - x_i \phi) + E(\dot{p})$$

$$= \sum p_i x_j f_j q_{ji} - \bar{p}\phi + E(\dot{p})$$

$$= \sum\nolimits_{ij} p_j x_j f_j q_{ji} - \bar{p}\phi + \sum\nolimits_{ij} (p_i - p_j) x_j f_j q_{ji} + E(\dot{p})$$

$\sum_{ij}(p_i - p_j)x_j f_j q_{ji}$ is equal to $E(f\Delta_m p)$ (the additional part in the expanded Price equation). $\sum_i q_{ji} = 1$, therefore

$$\dot{E}(p) = Cov(f, p) + E(\dot{p}) + E(f\Delta_m p) \tag{2.8}$$

Equation 2.8 is the same as equation 2.7. It has therefore been shown that the replicator-mutator equation and the expanded Price equation are equivalent.

It can also be shown that the replicator equation is equivalent to the standard Price equation. It is known that

$$\dot{E}(p) = \sum\nolimits_i p_i \dot{x}_i + E(\dot{p})$$

$\dot{x}_i$ can be replaced with the replicator equation (equation 2.3):

$$\dot{E}(p) = \sum\nolimits_i p_i (x_i[f_i(X) - \phi]) + E(\dot{p})$$

It is known that $\sum_i x_i f_i p_i - \bar{p\phi} = Cov(f, p)$. The equation can therefore be rewritten as

$$\dot{E}(p) = Cov(f, p) + E(\dot{p}) \tag{2.9}$$

Equation 2.9 is the same as equation 2.6, the standard Price equation. It has therefore been shown that the replicator equation and the standard Price equation are equivalent. $Cov(f, p)$ is the covariance of trait p with fitness f and therefore describes selection. $E(\dot{p})$ is the expected change in the values of trait p (which may be caused

by factors such as environmental change). $E(f\Delta_m p)$ describes mutation among types i and j [53].

It should be noted that, in Page and Nowak [53], the Price equation does not include mutation whereas the expanded Price equation does. However in Price's original paper, there was no mention of mutation at all. This is due to a difference in labelling. In Page and Nowak [53], $x_i(t)$ is said to denote "the relative abundance of type i individuals at time t." However Price's definition is that $x_i(t)$ denotes "the relative abundance of individuals at time t that are derived from type i individuals at time 0." Due to this, the Price equation may sometimes be referred to as the replicator Price equation, and the expanded Price equation as the replicator-mutator Price equation, when using the definition given by Page and Nowak. If labelling originally given by Price is used, the equation should remain as the Price equation.

2.2 Population Genetics

The previous section considered some of the principles and equations that are used to study the dynamics of evolving populations in terms of mutation, fitness and selection. The following section considers how changing evolutionary dynamics can influence the frequency of alleles.

Population genetics aims to understand the effects of evolution on populations of individuals. It comprises Darwin's theory of adaptation by natural selection [46], and Mendel's theory of inheritance. There is consideration of the change in the frequency of genes, as well as prediction of the effect of natural selection. The extent to which natural selection can influence the genetic variation within a population depends upon the amount of variation already present; if there is little variation, there will be fewer genes to select from. Factors that can influence variation include mutation, selection,

and random genetic drift.

2.2.1 Random Genetic Drift

Random genetic drift is the term used to describe the changes in allele frequency that occur by chance [54, 55]. Large populations are more likely to contain representatives from across the entire gamete pool (i.e., the set of all available gametes), whereas smaller populations are more likely to contain only a fraction. Consequently, in small populations, alleles will often be present at much higher or lower frequencies than they would be in larger populations. Mendel's theory of inheritance is a fundamental concept that underlies random genetic drift. The theory is comprised of two laws: the Law of Segregation and the Law of Independent Assortment. These laws state that, in a diploid organism, the two copies of each gene are separated so that each gamete receives only one copy, and that the assortment of alleles of different genes occurs entirely independently. Allele frequencies will therefore change randomly each generation.

If the number of gametes in a sample is represented as $2N$, the probability that i alleles of type A are present in that sample will be:

$$\binom{2N}{i} p^i q^{2N-i}$$

This is a binomial probability where $\binom{2N}{i}$ means $(2N)!/i!(2N-i)!$, and p and q are the frequencies of the alleles A and a within the set of gametes being sampled. The sum of frequencies p and q must always be 1. The value of i must be between 0 and $2N$. If there is random genetic drift, the new frequency of A alleles will be

$$p' = \frac{i}{2N}$$

This is simply the number of A alleles divided by the total number of gametes in the sample. As the alleles must be either A or a,

$$q' = 1 - p'$$

Allele frequencies are very difficult to predict in a population, as they change sporadically due to random genetic drift. However it is possible to predict average behaviour. When modelling random genetic drift, the following assumptions are made:

- The organism is diploid (i.e., has two copies of each chromosome).
- Reproduction is sexual (as opposed to clonal).
- The generations do not overlap.
- The population is made up of many distinct subpopulations (which arise from a large initial population). Each has a constant size denoted by N.
- Mating occurs randomly within each subpopulation, but not between.
- There is no mutation or selection.
- In addition to the above assumptions (which constitute the Hardy-Weinberg model [54]), we also assume there is an equal proportion of males to females, and each individual has an equal fitness.

Table 2.1: **Frequencies of the possible genotypes according to the Hardy-Weinberg law.**

Genotype	**Frequency**
AA	p_i^2
Aa	$2p_iq_i$
aa	q_i^2

This model does not accurately reflect a real biological system, as in reality many of the assumptions will not be true; it is an 'ideal' population. However it does provide a point of comparison, acting as a standard to which non-ideal (real) populations can be compared. The ideal population can be thought of either at the level of the whole population or at the level of an individual subpopulation. In a subpopulation i of an ideal population, the frequency of alleles A and a will be p_i and q_i respectively. The frequencies of the possible genotypes can be specified according to the Hardy-Weinberg law (Table 2.1).

A population may consist of multiple finite-sized subpopulations. A subpopulation is a group of individuals typically distinct from the rest of the population geographically, with little migration occurring between different subpopulations. In the total ideal population (that consisting of all subpopulations), there may be a decrease in the overall proportion of heterozygous individuals if enough time has passed so that each finite-sized subpopulation drifts towards either the A allele or the a allele. However, if the subpopulations are very large, random genetic drift may be of such negligible importance that we still see Hardy-Weinberg equilibrium. Returning to the case where random genetic drift is important, the lack of heterozygosity eventually observed is in effect the outcome that would also be seen due to inbreeding of organisms.

Alleles that are both descended from and identical to an ancestral allele can be described as being *identical by descent* [25, 54]. Let F_t be the probability of two randomly selected alleles (from one subpopulation) both being identical by descent. Assuming each subpopulation is of fixed size, and gametes combine completely randomly, any two alleles from within a subpopulation have the potential to be identical by descent.

Consider a population in which there are $2N$ alleles at generation $t-1$. There are n unique alleles, $\alpha_1, ...\alpha_n$, each of which makes up $1/2N$ of the population. Randomly select two alleles from generation t. The probability that both alleles are the same ($\alpha_i\alpha_i$) is simply the frequency of allele α_i $=1/2N$. In this case, the probability of identity by descent is equal to 1, as identical alleles must be descended from the same unique parent in this example. Conversely, the probability that the alleles are different ($\alpha_i\alpha_j$) is $1-1/2N$. Denote the probability of identity by descent $F_{(t-1)}$. The probability of two randomly selected alleles being identical by descent can therefore be given as:

$$F_t = 1 - \left(1 - \frac{1}{2N}\right)^t \tag{2.10}$$

Equation 2.10 is used as an approximation for the cases of both sexual and asexual reproduction. In the case of sexual reproduction, an assumption is made that half the gametes must be from males and the other half from females. This could be represented in equation 2.10 by replacing N with $N+1/2$. However, because $1/2(N+1/2)$ and $1/2N$ will be very close in value, equation 2.10 is usually considered as a good approximation.

Alleles that are identical by descent must also be homozygous. This is known as *autozygosity*, and F_t becomes the probability of *autozygosity* in an individual from generation t. Consequently, $1-F_t$ is the probability of *allozygosity* (when two homozygous alleles are not identical by descent, but are instead unrelated). F_t will eventually reach

a value of 1 as each subpopulation becomes fixed for either allele A or allele a, where fixation is the process by which each subpopulation becomes made up entirely of one type of allele due to a combination of random genetic drift and selection [54].

Consider a diploid population with equal numbers of males and females; the actual population size can be represented as N_a and the effective population size as N_e, where the effective population size refers to the number of individuals in the population that contribute offspring to the next generation. If a novel mutation occurs, its initial frequency will be $p = 1/2N_a$. Assuming the mutation is selectively neutral, the probability of eventual fixation will be $1/2N_a$, while the probability it will become lost will be $1 - 1/2N_a$. In general, fixation of a new allele is a long process, taking on average $4N_e$ generations. Conversely, a new allele may be lost very quickly, taking on average $2(N_e/N_a)$ generations [54].

2.2.2 Maintaining Alleles in Populations

The main source of genetic variation is mutation. The rate of spontaneous mutation in higher eukaryotes is usually small, at around 10^{-4} to 10^{-6} mutations per locus per generation. This means there is little chance of allele frequencies changing significantly within the space of a few generations. Conversely, over many generations, mutation can cause significant genetic change. Models have been developed that predict the effects of mutation on genetic variation [54]:

- The *infinite-alleles model* assumes that for each mutation, a novel allele is produced (in a finite population). It provides a good starting point to which other models can be compared.

- The *stepwise-mutation model* assumes that for each mutation, the resulting protein is altered to an extent that, during electrophoresis, it will migrate at a rate

increased or decreased by one unit. Electrophoresis separates proteins by applying an electric charge across a gel. This causes the proteins to move across the gel based on their size.

For the infinite-alleles model, the number of alleles required to be present so that loss of alleles by random genetic drift is exactly balanced by the generation of new alleles by mutation represents the number of alleles that can be maintained in the population. The amount of differentiation in the population is represented by the *fixation index*, F_ST. Fixation index represents the difference between genetic sequences (genetic polymorphisms), and is related to identity by descent in that it measures how related two individuals from a subpopulation are in relation to the total population.

Let N represent the population number, and μ the mutation rate. In the infinite-alleles model, each allele that emerges must be novel, i.e., an allele can only ever emerge once. Let there be n alleles, each with frequency p (i.e., $p_1, p_2, \ldots, p_n$). The proportion of homozygous individuals in terms of allele frequency is therefore represented by $\sum p_i^2$. Homozygosity can also be expressed in terms of fixation index. In the infinite-alleles model, all homozygous individuals must also be autozygous, in that both of their alleles must have descended from the same ancestral allele. Homozygosity is therefore equal to the fixation index. In the case of autozygous individuals, both alleles are identical to each other, as well as being identical by descent. If it is assumed that there has been no mutation in the time of one generation, the probability of no mutation of either allele can be included, represented by $(1-\mu)^2$. The number of alleles that can be maintained in the population, the equilibrium value of F_t, is given in [54] as:

$$\hat{\mathbf{F}} = \frac{1}{4N\mu + 1} \tag{2.11}$$

This means that mutation will increase the frequency of alleles in the population (that are not contributing to selection), until the value of $\hat{\mathbf{F}}$ is reached. N can be replaced in Equation 2.11 with the effective population size N_e, to consider only those individuals contributing offspring to the next generation. In population genetics, $4N_e\mu$ can be written as θ, therefore

$$\hat{\mathbf{F}} = \frac{1}{1+\theta} = \frac{1}{4N_e\mu + 1}$$

2.2.3 Natural Selection and Fitness

Random genetic drift promotes the genetic divergence of a subpopulation of a species, while migration slows the rate of divergence due to organisms moving between subpopulations. The effect of a mutation on genetic variation will be dependent on the balance between the processes of migration, random genetic drift, and natural selection; the dispersion, loss, or maintenance of a mutation depends upon the structure and processes occurring in a population [54]. The process of natural selection occurs at the level of the phenotype as opposed to the genotype and is dependent on multiple genetic loci and environmental factors which have an effect on the phenotype of an individual. The type of effect is determined by the type of selection. There are three modes of selection, namely *directional*, *stabilising*, and *disruptive* selection. Consider the range of all possible phenotypes. If phenotypes at one extreme of the range are favoured, selection is said to be directional. If selection favours phenotypes at both extremes of the range, this is classed as disruptive, while selection of phenotypes in the middle of the range is described as stabilising.

While the process of selection occurs at the level of the phenotype, the consequences of selection can be observed by studying changes in allele frequency at a given locus.

Natural selection can be quantified by using the idea of fitness. In population genetics: "The fitness of a genotype is the average number of offspring produced by individuals of that genotype"[1]. Aspects of fitness include both viability and fertility (or fecundity), which refer to the probability of survival to reproductive age, and average number of subsequent offspring respectively. Half of each offspring is considered to belong to each parent. The precise definition of fitness has been subject to debate [56]. For example, Waddington defines the fittest individuals as being those that are: "most effective in leaving gametes to the next generation" [57, 56]. Dennett defined fitness in terms of individuals x and y, stating that : "x is fitter than y if and only if x's traits enable it to solve the 'design-problems' set by the environment more fully than y's traits do" [58, 56]. Fitness has also been treated as a probability and thus the definition becomes: "x is fitter than y in $E = x$ has a probabilistic propensity > 0.5 to leave more offspring than y" [56]. This may not always be true, as there are many external factors that can affect the number of offspring such as availability of food. Rosenberg and Bouchard therefore develop the definition: "x is fitter than y = probably x will have more offspring than y, unless their average numbers of offspring are equal and the temporal and/or spatial variance in y's offspring numbers is greater than the variance in x's, or the average numbers of x's offspring are lower than y's, but the difference in offspring variance is large enough to counterbalance y's greater number of offspring" [56]. The variation in interpretation of fitness means the precise definition may vary depending on the circumstances of study.

Fitness can be measured as absolute or relative. If there is a genotype AA, of which 3/4 survived to reproductive age, and 6/3 went on to produce offspring, the fitness of AA will be 3/4 multiplied by 6/3, i.e., 1.5. This is the *absolute fitness*. The

[1]Note the difference in the population genetics definition of fitness compared with the definition of a fitness value given in section 2.1.

ratios of absolute fitnesses give *relative fitnesses*. The relative fitnesses of genotypes $AA : Aa : aa$ can be denoted $\omega_{11} : \omega_{12} : \omega_{22}$.

The term fitness may be used to describe the viability, with fertility omitted for simplicity. It should be noted that fitness is a property of an entire genotype, not a single locus. Fitness is also dependent on the environment, and is consequently rarely constant; it can differ between different subpopulations and between different generations, and will vary with environmental change [54].

2.2.4 Dominance

In general, an individual with two different alleles will display the phenotype of the dominant allele. In the majority of cases, mutant alleles are recessive while the non-mutant wild-type alleles are dominant [45]. In diploid organisms, loss-of-function mutations at one of the two alleles often have no observable effect: this is due to dominance. Dominance increases genetic robustness by masking the presence of negative mutations. In some cases, two alleles may be different but equally dominant and so both visible in the phenotype; this is co-dominance. Co-dominance can be seen in human blood types, for which there are three alleles encoded at one locus [59]. Alleles A and B are co-dominant, while both are dominant to O.

Fisher [60] proposed that the evolution of genetic elements (known as dominance modifiers) that control dominance interactions between alleles is the cause of such dominance relationships. Wright did not dismiss the possibility of such elements existing, but showed that, at least in the case of recessive deleterious mutations, there would be little selection pressure to evolve them [61]. Fisher's theory was also claimed to be falsified by the fact that, even in *Chlamydomonas*, which spends the majority of its life cycle as a haploid, the wild-type alleles are dominant; selection on heterozygotes

would be inefficient as they can exist only in the short diploid stage of the life cycle, suggesting dominance evolved by some other mechanism than by Fisher's dominance modifier theory. Over time, Wright's physiological theory (see [62]) has become the widely accepted theory of the evolution of dominance.

In *Brassica*, dominance is controlled by small RNAs binding to the promoter region of recessive alleles when there is a dominant allele also present [61]. This silences the recessive allele by mediating methylation of the promoter region, and leads to expression only of the dominant allele. This observation means that a dominance modifier has been identified which corresponds exactly to the type of dominance modifier originally hypothesized by Fisher in 1928. Wright's physiological explanation would be ineffective in this case because this small RNA specifically controls the interaction between alleles in heterozygotes without affecting the expression level in homozygotes. This new finding confirms that Wright's explanation does not apply to all observed cases of dominance. Although the mechanism described might be considered to be a special case, Billiard and Castric [61] argue that the possibility that such dominance modifiers exist can be viewed as validating Fisher's theory. Wright's theory only addresses a specific category of genes: enzyme-encoding genes involved in metabolic networks. Genes not in this category, e.g., structural genes or genes involved in receptor-ligand interactions, still lack a general theory of dominance. Also, Wright's primary argument for dismissing Fisher's theory revolved around the weakness of natural selection for dominance when heterozygosity is low, and therefore does not cover cases where selection is strong or heterozygosity is high. Wright himself acknowledged the possibility that dominance modifiers *per se* could evolve in nature as 'special cases'.

Kacser and Burns [63, 64] proposed an explanation of dominance in terms of genes for metabolic enzymes (which make up a considerable fraction of the genome).

Metabolic enzymes are part of pathways, or larger networks, in which the substrates of any one enzyme are the products of other enzymes. Mutations that occur in metabolic genes affect the action of enzymes, and consequently the efficiency of chemical reactions. The amount of substrate converted into product per unit time is known as the flux through a reaction. Some metabolic fluxes are highly correlated with fitness, for example, for reactions that generate essential amino acids, or whose products have protective functions such as pigments. Dominance in metabolic genes can therefore be expressed in terms of flux. The question is: how does the flux F through some critical chemical reaction (and therefore fitness) change if the activity E_i of enzyme i somewhere else in the system changes (by a given amount)? This change is described by the flux control coefficient (which can be represented as a fractional change):

$$C_i = \frac{\triangle F/F}{\triangle E_i/E_i}$$

In a system that is robust, a large change in enzyme activity will only change an important flux by a small amount. If flux is insensitive to a change in an enzyme's activity of 50% or more, the gene encoding the enzyme will be dominant to a version of the gene which contains loss-of-function mutations. The flux control coefficient C will depend upon the magnitude of the change $\triangle E_i$.

In Mendel's notation, A is used to mean the dominant allele, while a is used to mean the recessive allele [65]. H can be used instead of Aa to represent some hybrid of these two alleles. Alleles are not themselves dominant or recessive; they are either normal or abnormal (mutant). In terms of metabolic pathways, a dominant trait A corresponds to the normal (100%) activity of an enzyme. When mutated, the activity levels of the enzyme decrease. A hybrid H corresponds to any flux level that can not be distinguished from its parent, even if the activity levels of the enzyme have been

significantly reduced. In hybrids, significant reduction in enzyme activity levels does not have much of an effect on flux (where the amount of substrate converted into product per unit time is known as the flux through a reaction; the decrease in activity of enzymes in a pathway that does not correspond with a decrease in flux, will not display a change in phenotype and so the wild-type phenotype will be observed despite the presence of mutation). It is not that individuals with genotype Aa only display trait A, but rather the hybrid trait H is indistinguishable from trait A. It may be the case that while the hybrid trait remains outwardly indistinguishable from A, the relative concentration of products of the pathway has changed with the mutation in enzyme activity. In some cases, such as when A and a refer to different colours, the hybrid trait H may be distinguishable as a blend of the two colours. a refers to an enzyme activity approaching, but not necessarily reaching, 0% activity [65].

In this chapter the basic concepts of evolution and adaptation have been introduced, along with key ideas and equations in the broader areas of evolutionary dynamics and population genetics. The following chapter introduces the different types of mutation, the nature of adaptation, optimal mutation rates and error thresholds, and the concept of mutational robustness and survival-of-the-flattest.

Chapter 3

Mutation, Variation and Adaptation

3.1 Mutation

The previous chapter examined how changing evolutionary dynamics can influence whole populations of individuals. The following chapter considers in detail one of the key parameters of evolution; the introduction of variation through mutation. Many mutations have a detrimental effect, but those that are beneficial enable adaptation to occur through selection. Adaptation allows individuals to survive to reproduce, increasing their fitness and ensuring as many offspring as possible will share their genotype. However, not all mutations will lead to a change in an individual's phenotype. Such neutral mutations mean that an individual that has not changed in fitness may still be undergoing significant genetic change. Introduction of genetic variation is key to evolution, meaning the mutation rate can have a significant effect. More precisely, the optimal mutation rate, critical mutation rate and error threshold determine how efficiently adaptation will occur.

3.1.1 Beneficial Mutations

A great deal of the work done in the field of population genetics has concentrated on the effect of deleterious or neutral mutations, i.e., mutations which cause the loss of genetic material, or have no effect on fitness respectively. Such mutations are more common than those that have a beneficial effect on the fitness of an individual. However, beneficial mutations are those which allow a population to undergo adaptation. Haldane [66] considered the theoretical effect of beneficial mutation [18]. His aim was to determine whether natural selection was sufficient to overcome the effect of Mendel's law of segregation. Consider a unique mutation which has a beneficial effect in heterozygote individuals. Let the beneficial effect be represented by s. If a heterozygote has a beneficial mutation, it will have an increased fitness of $1+s$. However, the law of segregation means that the beneficial mutation may not be passed on to the offspring, with the other allele being passed on instead.

Haldane used branching process theory to determine the probability of a beneficial mutation becoming fixed in a population. Branching process theory is a way of modelling a population, in which individuals in generation n reproduce to give individuals in generation $n+1$, with some fixed probability. Haldane determined that fixation of beneficial mutations (p_{fix}) occurs around $2s$ of the time. To give an example, if a new mutation increases fitness by $s = 0.001$, then $p_{fix} = 0.002$. This new mutation must occur approximately $1/p_{fix} = 500$ times in order for natural selection to fix it in the population [18]. This fixation will occur with 63% likelihood. The probability that the new mutation is not fixed is $1 - p_{fix}$. The probability that the new mutation occurs $1/p_{fix} = 500$ times and does not get fixed is $(1 - p_{fix})^{(\frac{1}{p_{fix}})}$. The probability that the mutation gets fixed in this example is therefore $1 - (1 - p_{fix})^{(\frac{1}{p_{fix}})} = 0.63$.

3.1.1.1 The Distribution of Fitness Effects Among Beneficial Mutations

Beneficial mutations are much rarer than deleterious or neutral mutations, therefore they can be difficult to study. The wild-type allele usually has a very high fitness; very few mutations will be good enough to increase this fitness. Let N represent population size, s be a selection coefficient, and μ represent mutation rate. Consider a case where selection is strong ($|Ns| > 1$), and mutation is weak ($N\mu \ll 1$); almost all of the population will be wild-type. The wild-type can mutate to a number of different sequences. Each of these possible sequences will have a certain fitness value, although the distribution of fitness will not be known. The fitness value of the wild-type must be higher than that of any of the mutant m sequences (otherwise it would not be present in such a high proportion of the population). If a beneficial mutation was to arise, its fitness value would have to be even higher than that of the wild-type, either due to a fitter sequence arising by chance, or by a change in the environment causing a drop in the fitness of the wild-type [67, 18].

Gillespie– [68, 18] wanted to know how great the difference in fitness was between wild-type individuals and beneficial mutants. He made the assumption that there could only ever be one beneficial mutation, an assumption that was later removed by Orr [18]. It was shown that the gap in fitness between wild-type and beneficial sequences is exponentially distributed; this gap represents the effect of a beneficial mutation on fitness. This was shown to be independent of the overall fitness distribution across the population. These results suggest that it should be much more likely for mutation to produce an allele which only increases fitness slightly; significantly fitter alleles will arise much more rarely.

The theory behind the distribution of fitness effects of beneficial mutations has been studied experimentally. Kassen and Bataillon [69] used the bacterium *Pseudomonas*

fluorescens, and studied mutants that had been derived before forces such as selection, genetic drift, and clonal competition could have an effect, i.e., mutants that had arisen naturally during population expansion. Strains were used that had acquired a single mutation, giving them resistance to the antibiotic nalidixic acid. Fitness of the bacteria was determined initially in an environment without any antibiotic present. These nalidixic acid-resistant mutants were then isolated and their fitness was determined in both the presence and absence of antibiotics (selective and permissive environments respectively). A large number of mutants were studied due to the rarity of beneficial mutations. As all the mutants being studied were resistant to nalidixic acid, they were all classed as beneficial mutants when placed in the presence of the antibiotic. The distribution of their absolute fitness could be seen to be roughly normal, and analysis of variance showed there to be significant genetic variation in fitness among the population. In the environment with no antibiotics, the majority of the mutants could be classed as deleterious in comparison with the wild-type. The distribution of fitness was shown to be superficially the same as that when antibiotics were present, and similarly there was shown to be significant genetic variation in fitness [69].

The second experiment tested the prediction that the distribution of fitness effects will always be exponential regardless of the fitness of the wild-type, i.e., the distribution of beneficial effects is invariant. This involved looking at a number of the highest-ranking mutants from the initial population of mutants in the environment with no antibiotics. Their fitness was assayed in four qualitatively different environments, and the relative fitness of the wild-type was seen to vary between them. However, the distribution of fitness effects was exponential across all four environments, and the mean fitness effect statistically the same for each. This is an important result because it contradicts theory slightly. Orr [67] suggested that the mean fitness effect should

be dependent on the size of the gap in fitness between the fittest and the next-fittest mutation. This fitness gap is likely to vary depending on the environment. This is contradictory to the results obtained by Kassen and Bataillon which suggest the mean fitness effect is conserved across environments. Another observation by Kassen and Bataillon is that when genotypes are ranked from fittest to least fit, the size of the fitness gaps between genotypes of neighbouring rank increases with absolute fitness, and therefore increases with fitness rank; this is consistent with what would be expected if the distribution of fitness effects was exponential. The authors conclude that the distribution of fitness effects among beneficial mutants can be characterised by an exponential distribution, and that the distribution of fitness effects is invariant. This further suggests that the beneficial mutations that become subject to natural selection can be characterised by many mutations of small effect, and few mutations of large effect [69].

While experiments by Kassen and Bataillon have shown the distribution of beneficial effects to be exponential, others have shown them to be closer to gamma distributions. This may be explained by the fact that, in some circumstances, an exponential distribution can resemble a gamma distribution as can be seen in Figure 3.1. Sanjuán [70] showed the distribution of beneficial effects among new mutations in vesicular stomatitis virus to be statistically closer to the gamma distribution than the exponential. Rokyta [71] allowed bacteriophages to adapt to a specific host bacterial species, and new mutations were isolated and sequenced. They showed that the distribution of beneficial effects was approximately uniform, meaning there was an equal chance of a beneficial mutation having either a large or small effect. This contradicts the results of the microbial experiments which, like the theory, saw that there was a much greater chance of getting a beneficial mutation which had a small effect. Orr [18] speculates

that the differences in experimental results may be due to the difficulty in characterising new mutations, and the rarity of beneficial mutations. It may also be explained with Fisher's Geometric Model of Adaptation, which states that when there are only a few characters, the distribution of fitness effects among beneficial mutations will be close to uniform, whereas when there are many characters, it will be closer to exponential (see section 3.1.2). In the context of the experiments, bacteriophages have relatively short sequences, meaning there is a greater chance of a beneficial mutation having a large effect. Conversely, as the sequences get longer, any beneficial mutation is more likely to have a smaller effect, as there will be many more characters for it to affect [18].

3.1.2 The Nature of Adaptation

An organism is said to be adapted to its environment if, in the case of the environment being changed in any way, the organism would become less well adapted [45]. In other words, an organism that is highly adapted to a specific environment will likely be disadvantaged by environmental change. According to Fisher's Fundamental Theorem of Natural Selection, "the rate of increase in fitness of any organism at any time is equal to its genetic variance in fitness at that time" [45, 72]. This means that natural selection will tend to lead to an increase in fitness. Genetic variance refers to the additive genetic variance, which is the variance based on the average effect on fitness of substituting one allele for another.

Fisher states that the statistics of any situation where one thing is made to conform to another can be illustrated geometrically. This is known as the Geometric Model of Adaptation (Figure 3.2) [45, 73, 74]. Consider a fixed point O, and a point A which can move. The degree of conformity can be represented by how closely point A approaches

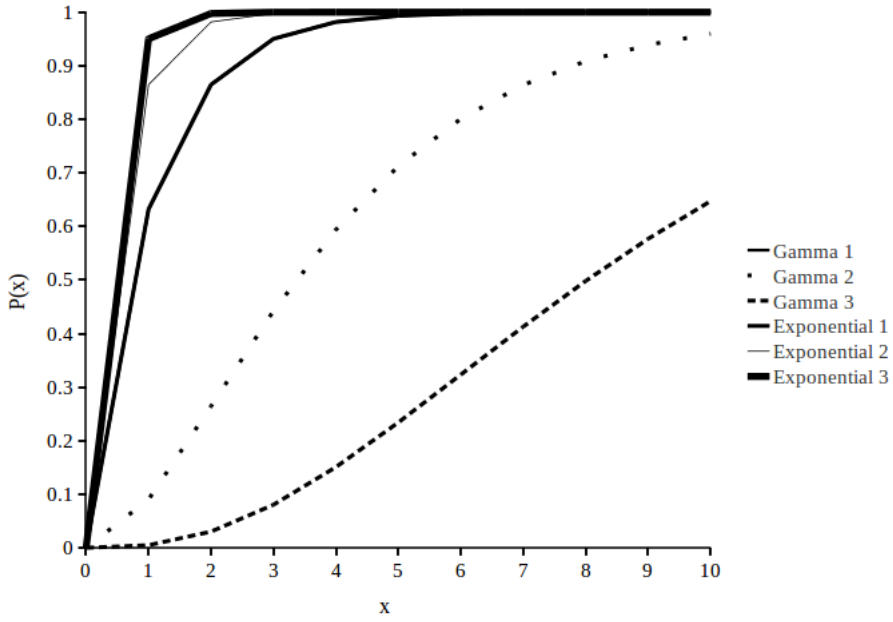

Figure 3.1: **The gamma distribution versus the exponential distribution** The gamma distribution ($F(x;k) = x^{(k-1)}e^{\theta}/\Gamma(k)$) and the exponential distribution ($F(x;\lambda) = 1 - e^{-\lambda x}$). x is varied from 0 up to 10 while the other parameters are kept constant at the following values:

Gamma 1: $k = 1$, $\theta = 1$	Exponential 1: $\lambda = 1$
Gamma 2: $k = 2$, $\theta = 2$	Exponential 2: $\lambda = 2$
Gamma 3: $k = 3$, $\theta = 3$	Exponential 3: $\lambda = 3$

It should be noted that Gamma 1 and Exponential 1 match exactly, demonstrating the potential similarity between the gamma and exponential distributions.

point O. There is a sphere which is centred at point O and passes through point A. The inside of the sphere represents positions that are better adapted than point A, while the outside of the sphere represents positions that are worse adapted. For example, allow point A to move by a fixed distance r (in any direction); if it were to end up inside the sphere, adaptation would have improved. Conversely, if it were to end up outside the sphere, adaptation would have been impaired.

If r is very small, the chance of improvement or impairment occurring is approximately equal; the chance of improvement tends towards $\frac{1}{2}$ as r tends towards zero. However, if r is as big as the diameter of the sphere (or greater), the chance of improvement is zero, as all points within the sphere are less than this distance away. We can therefore say that if $r \geq d$, chance of improvement is zero (where d is the diameter of the sphere), while if $r \approx 0$, chance of improvement is $\frac{1}{2}$. The latter is the maximum probability of improvement (the limiting value). For any value of r between these limits, the probability of improvement is approximately

$$\frac{1}{2}\left(1 - \frac{r}{d}\right)$$

This means there is a steady decrease in the chance of improvement from $\frac{1}{2}$ (when $r = 0$), to 0 (when $r = d$). Point A may represent either the organism or the environment; a very small change in either has an almost equal chance of improving or impairing adaptation. As the magnitude of the change increases, the chance of improvement decreases until it reaches zero (or at least negligible).

It is possible to extend this model to include more than three dimensions. Increasing the number of dimensions changes the way the probability of improvement changes with respect to the magnitude of r. There will still be a limiting value of $\frac{1}{2}$, with the probability of improvement decreasing from $\frac{1}{2}$ to 0. If the number of dimensions is

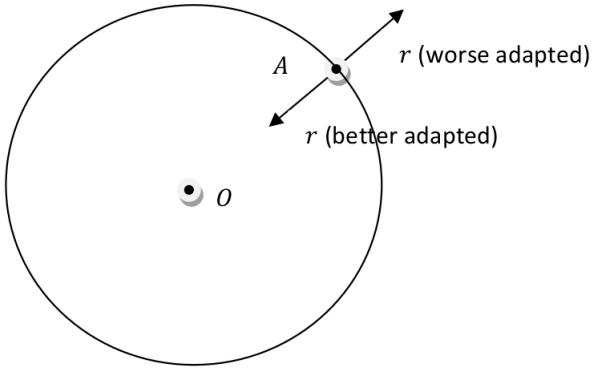

Figure 3.2: **Fisher's Geometric Model of Adaptation**. In a sphere centred at O and passing through point A, mutation can move point A by a fixed distance r in any direction. The direction determines adaptation.

denoted n, the standard magnitude of change is represented by $\frac{d}{\sqrt{n}}$, where 'change' refers to the fixed distance r that point A moves (as mentioned previously). The probability of improvement is determined by the ratio of the change being considered to the magnitude of the change, i.e., $r : \frac{d}{\sqrt{n}}$. In other words, the higher the adaptation, the smaller the value of both $\frac{d}{\sqrt{n}}$ and the probability of improvement. Consider changes of a given magnitude occurring at random in all directions. There are two opposite selective agencies. One selects for all changes which increase adaptation, while the other selects for all changes which decrease adaptation. Very small changes are equally likely to be affected by either selective agency, while larger changes are more likely to be negative [45].

Gillespie's 'mutational landscape' model [68], follows adaptation at the level of a gene or small genome that is L base pairs long [75]. Point mutation rate is assumed to be small ($N\mu << 1$) and selection strong relative to population size ($Ns > 1$, where s is a selection coefficient). Under these assumptions, at any one time the population

will be fixed for the wild-type (assumed to be the fittest) allele. Imagine there is an environmental change introducing the potential for at least one beneficial mutation. Each of the L positions in the wild-type sequence can mutate to 3 other possible bases. The number of possible sequences that can be mutated to by changing only one of these L positions is therefore represented by $m = 3L$ [76, 68, 75]. Due to the rarity of double and triple mutants, these can be ignored [68]. This means that, under the mutational model, natural selection can be seen to focus on a single mutational step in sequence space. Even though the wild-type may have lost its position as the fittest allele, it will still have a high fitness, as it is unlikely that the wild-type will experience a sudden drop in relative fitness as the environment changes [75]. Let the wild-type have fitness rank i, where i is small. There are m alleles that can be reached with a single point mutation, and therefore $m+1$ relevant alleles including the wild-type itself. Assuming it takes a single step to better the wild-type, there will be $i-1$ single-step mutations, m, that will be beneficial. Eventually, one of these $i-1$ favourable alleles will become fixed as recurrent mutation counteracts accidental loss. There will now be a new wild-type and the process will repeat until the population reaches a sequence that is fitter than all of its single-step mutational neighbours (a local optimum) [75].

Both Fisher and Gillespie's models have limitations [75]. For example, nothing can be said about adaptation from standing genetic variation using Fisher's model, as it considers only new mutations. In addition, Fisher's model only considers a single instance of adaptation to a new optimum, not long term where the optimum is continually changing and so adaptation is continually occurring. This also applies to Gillespie's model, which does not account for the situation where the rate of environmental change is greater than the rate of substitution. It should be noted, however, that this situation is unlikely in the case of microbial evolution and therefore the mutational model can

still be applied. Gillespie's model assumes that selection is strong. Neutral mutations have no effect on fitness and are therefore not subject to selection, meaning another limitation of Gillespie's model is that it cannot be applied to a system with neutrality.

3.1.3 Neutral Theory

In natural systems, the genotype encodes the molecules that are assembled to produce the components of the organism; the phenotype. There are complex interactions between molecules as the organism develops in its environment. However, in many artificial systems there is no development process. Instead there is a direct mapping between the genotype and phenotype. The phenotype is essentially encoded in the genotype, a property which does not always generalize to natural systems. In a natural system, there may be a many-to-one mapping of genotype to phenotype; more than one genotype can produce the same phenotype. This means that it is possible to have a mutation in the genotype that has no effect on the phenotype. This is the *neutrality hypothesis*, and such mutations are known as *neutral mutations* [17, 77]. It has been suggested that in natural systems only very few non neutral mutations are beneficial [78].

There are 64 codons in total (combinations of three bases) that code for 20 amino acids [51]. There is therefore a large amount of redundancy in the genetic code. A mutation may not change the amino acid being coded for, and will therefore be neutral. This can be taken one stage further, in that a change in amino acid may not have any effect on the structure (and therefore function) of the resultant protein. The function of a protein is dependent on the way the amino acid chain is folded; replacing one amino acid may have little to no effect on the protein structure, have no effect on the phenotype, and therefore be classified as a neutral mutation. However, some amino

acids are critical, and replacing them can completely change the protein structure; this will likely cause a change in protein function. Unless there is another protein present that can perform the same role as the mutant protein, there is likely to be an effect on the phenotype. Genes are often part of complex networks, meaning the mutant protein may well have played a role in the regulation of other genes, networks, or signalling pathways, having a potentially large effect on the phenotype [79].

In 1968, Kimura made the suggestion that in mammals, the majority of mutations have no effect on the phenotype of the individual, i.e., most mutations are neutral [17]. Based on the typical values in mammals for DNA chain length, and the number of base substitutions that translate to amino acid substitutions, he worked out that there has been approximately one base pair substitution in the mammalian population every 2 years, but that this usually does not translate to an amino acid substitution. At the molecular level such neutral mutation should be seen as the normal, while beneficial or deleterious mutation is the exception, in contradiction to population genetics which assumes homogeneity. The homogeneity assumption states that the mutation rate is so low that a new allele will become fixed in the population before the next mutation occurs; evolution is the movement of a homogeneous population throughout the fitness landscape. Neutral theory states mutation can cause changes in the genotype without there being any change in phenotype, and consequently *without* any movement across the phenotypic fitness landscape [79].

Neutral mutations may not directly alter a particular trait, but they may be involved in the regulation of other proteins. Some genes are part of complex regulatory networks and mutation of them may lead to the inhibition of certain proteins, consequently indirectly affecting a seemingly unrelated trait. Manrubia and Cuesta [79] suggest that there is virtually no single trait in multicellular animals or plants which is not

dependent on a combination of genes acting together. The phenotype should not be considered as the linear combination of traits, but rather the effect of the genome as a whole. It is possible to have two individuals which appear to have very similar phenotypes, but with very different genotypes. This means that a mutation could have a large effect on the phenotype in one individual, and have no effect in another. This contradicts the population genetics idea that there is an optimum genotype which has maximum fitness, and that mutation of this genotype will cause fitness to decrease. Consider the distribution of phenotypes in genotype space. Individuals with the same phenotype will have the same fitness, therefore it is possible for an individual to move in the space without any alteration in fitness. In such cases, it can be said that the mutant genotype belongs to the same *neutral network* as the parent genotype. There may be instances where a single point mutation is enough to cause an individual to move from one neutral network to another and change in phenotype, i.e., where two neutral networks are close to each other in genotype space [79]. Many real search problems have a genotype search space which breaks down into such neutral networks. This often leads to long periods of fitness stasis during which there is search along a neutral network, with occasional leaps in fitness as the search moves to a higher neutral network [80].

Schuster [28] discusses Motoo Kimura's contribution to work on neutral evolution. Neutral genotypes (with the same fitness values) will drift randomly in sequence space until one genotype becomes fixed. Kimura suggested that the average time to replace one genotype with another is equal to the reciprocal of the mutation rate, $\tau_{subst} = 1/p$. It is therefore completely independent of the population size. Conversely, the time it takes for a mutant to become fixed in the population is proportional to the population size, $\tau_{fix} = 4N_e$ (where $4N_e$ represents the effective population size) [28].

The notion of evolution exploring a neutral network silently until it finds a better phenotype, then switching and silently exploring this new neutral network, follows the notion of punctuated equilibrium. This is when there appears to be little evolutionary change for long periods, followed by rapid change which may lead to speciation. A real world example of this can be seen with the influenza virus; new populations of influenza emerge every few years, which is why it is necessary to develop new vaccines [79]. Another well-studied example of neutrality in the real world is that of RNA secondary structure formation. Studies have lead to the identification of four properties of RNA genotype to phenotype (or RNA sequence to shape space) mapping [81, 82, 83, 84]:

- There are more sequences than there are structures (i.e., many-to-one genotype to phenotype mapping).
- There are only a few common structures, and many rare structures; most of the sequences form one of a few discrete structures.
- The distribution of RNA sequences that map to the same structure seems to be random in sequence space.
- There are interconnected neutral networks in shape space. All parts of the shape space can be reached by the process of random neutral drift. This means the population will not get stuck on local optima; this is given the term *constant innovation*.

3.1.3.1 Neutrality in Multiple-peak Landscapes

Evolution is the process of climbing the peaks in a fitness landscape. However, problems can arise when there is more than one peak, i.e., there are many local optima. Once an individual has reached a local optimum, it would need to climb down the current peak

in order to find the global optimum. This would mean reducing its fitness, which is contradictory to adaptation. The only way to get from a local optimum to the global optimum without reducing fitness would be through recombination with individuals at other local optima. The problem with this is that the population must be sufficiently diverse so that there are other individuals on other peaks. Such diversity is reduced by both selection and random genetic drift. In systems with neutrality, a change in the genotype may not necessarily lead to a change in fitness; genotype change does not automatically imply a change in phenotype. In the case of such neutral mutations, individuals may drift, allowing them to reach other peaks in the landscape. In this way, neutral mutations can prevent an individual from becoming stuck at a local optimum, without the need for a reduction in fitness [78].

In artificial systems, use of a genetically converged population risks the population becoming stuck at a local optimum. This becomes far less likely if neutrality is introduced, as populations can continually move through genotype space regardless of fitness [85, 78]. As in natural systems, without neutrality, escape from a local optimum may require a temporary reduction in fitness which would be counterintuitive [84]. In terms of problem solving, the dependence of one solution on the existence or state of another solution means there is a many-to-one mapping of genotypes to a particular solution; there are neutral networks in genotype space [85]. Barnett [86] suggests that in systems with neutral networks, the search for high-fitness genotypes is done through neutral mutation and mutation to higher neutral networks. In addition to this, Barnett [86] claims that an efficient strategy for evolutionary search involves the action of independent 'netcrawlers'. This means evolution occurs through the action of single individuals each replicating and mutating, followed by elimination of either the parent or offspring depending on which is fitter; the population can be considered as a group of

independent netcrawlers. Rather than the whole population climbing a peak, systems with neutrality allow the population to mutate along neutral networks and explore the genotype space with no effect on fitness [78]. This enables the population to encounter a greater number of phenotypes or solutions, increasing the chance of finding one with a higher fitness.

3.1.3.2 Neutral Theory and Selection

Alleles that increase the fitness of an individual will tend to increase in frequency until they replace the ancestral allele and become fixed in the population. This is known as *positive* or *directional* selection (refer to section 2.2.3). Alleles that decrease the fitness of an individual will tend to decrease in frequency, a process known as *negative* or *purifying* selection. In some cases, an allele may be beneficial only in individuals that are heterozygous, in which case the allele will tend to remain in the population with an intermediate frequency; this is known as *balancing* selection. Natural selection is a directional process. This is in contrast to genetic drift which is a random process that changes allele frequencies due to the random selection of gametes that will go on to produce adult individuals in each generation [87].

Neutral mutations are not affected by natural selection. Natural selection works at the level of the phenotype, and neutral mutations are changes in the genotype that do not affect the phenotype. The frequency of alleles arising by neutral mutation will be entirely dependent on random genetic drift. The neutral theory claims that the majority of evolution at the molecular level is not due to selection, it is instead due to the random fixation of neutral (or nearly neutral) mutants due to sampling drift. Kimura's theory has been verified against actual sequence data. It was suggested that, if the majority of divergence of genetic sequences between species was due to

neutral evolution, there would be expected to be more changes in sequences that are functionally less important. Indeed, studies of sequence data showed that:

- In protein sequences, there is a higher rate of substitution among amino acids that have similar biochemical properties; such substitutions are not likely to affect protein function.
- Synonymous base substitutions (which do not cause a change in amino acid) occur at a higher rate than nonsynonymous substitutions.
- The non-coding parts of a gene (introns) evolve at a high rate.
- Pseudogenes (segments of DNA that look like genes but do not get transcribed) evolve at a high rate.

This is consistent with the neutral theory, but contradicts selectionist theory. Selectionist theory suggests that, if most substitutions were adaptive, there would be expected to be fewer changes in sequences that are functionally less important; the majority of changes would be in the coding parts of a gene, which is not the case. The difference between the neutral theory and the selectionist theory is the relative proportions of neutral and advantageous mutations that lead to sequence divergence and polymorphism. It is now accepted that while these proportions can vary between different species, a significant proportion of substitutions will usually be neutral and therefore neutral substitutions should not be ignored [87].

Mutations can range from highly deleterious to weakly deleterious, nearly neutral, neutral, weakly advantageous, to highly advantageous. The effect of selection on a mutation is dependent on the selection coefficient s (the fitness effect of the mutation), and on the effective population size N_e (the number of individuals in an 'ideal' population contributing offspring to the next generation). When $N_e s \ll 1$, the effect

of mutation will be determined by random genetic drift alone; when the population is small, random genetic drift will outweigh the effect of natural selection. In such cases, all mutations can effectively be considered to be neutral. This suggests that the proportion of neutral mutations can be expected to vary inversely with effective population size [87].

Another prediction of the neutral theory is that at selectively neutral sites (where there is no selection), the rate of substitution, where one allele is completely replaced by another within a population or species over time, is equal to the rate of mutation [17, 87]. If the rate of mutation per generation at a selectively neutral site is represented by u, and the population size is represented by N (in a haploid population), there will be Nu mutations at this site per generation. As there is no selection, all genotypes will become fixed with the same probability. The probability that an allele or mutation will become fixed is equal to its relative frequency in the population. In a haploid population, a new mutation will have a relative frequency of $1/N$; the probability of fixation will therefore also equal $1/N$. The rate of substitution can be represented by K, and calculated per generation:

$$K = Nu \times \frac{1}{N} = u$$

The process of *biased gene conversion* (BGC) can also contribute to changes in allele frequency in sexual populations. BGC occurs during crossover, when one allele gets copied and pasted onto another allele, i.e., converting the allele into a copy of the donor allele. Such gene conversion is said to be biased if one allele has a greater probability of conversion than the other. If this is the case, there will be a higher frequency of donor alleles in the population compared with the converted allele. A downside of BGC is

that it can favour the fixation of deleterious alleles. In some eukaryotes, BGC seems to favour fixation of mutations that go from AT to GC, leading to the formation of GC-rich regions; these GC-rich regions will occur where there is a high crossover rate. In mammals, BGC can lead to the development of regions where the local substitution rate can be up to 20 times higher than that of the rest of the genome [87].

3.1.4 Optimal Mutation Rates and Error Thresholds

In a landscape with a single fitness peak, the quasispecies is able to maintain its position surrounding the top of the peak so long as the mutation rate does not exceed a critical mutation rate known as the *error threshold*. Above this threshold, there is an error catastrophe and the population delocalises across sequence space [26, 13, 14, 27, 11, 28, 29]. The concept of an error threshold was introduced in Eigen [31] and later in Nowak and Schuster [32] based on the quasispecies equation (Equation 2.1). It is essentially a limit on the amount of information that can be maintained in the system, and as such is linked to the *information threshold*. Consider a simple evolving system; the information threshold introduces a paradox. Imagine there is a high mutation rate which is giving the system a high level of diversity. To allow adaptation to occur, and therefore to increase the amount of information that can be maintained, the system needs to evolve a molecular mechanism to reduce the mutation rate. However, to encode such a mechanism, the system will require long sequences initially. In other words, to allow the system to evolve the ability to maintain longer sequences, it must already have long sequences in which to code the necessary molecular mechanism; to evolve complexity, the system needs complexity. This is known as *Eigen's Paradox* [10, 88].

The greater the sequence length of genotypes in the system, the greater the error

rate [89]. Schuster [28] suggested the error threshold could also be considered as the 'localisation threshold' of the quasispecies in sequence space; the value of the error threshold will determine the maximum value of the mutation rate and therefore the movement of the population. Flatter fitness landscapes therefore have a less obvious error threshold compared with single-peak landscapes, in that single-peak landscapes have an obvious point around which the population will cluster and form the quasispecies [28].

Selection and mutation provide two forces (or pressures) on the population, and they can be combined into one matrix ($w_{ji} = f_j q_{ji}$) (see [14], p. 35). Selection draws the population closer to the highest fitness, while mutation is usually assumed to have a deleterious effect due to which the population drifts away from the highest fitness. Generally, the population converges to a stable (equilibrium) state that is defined by an eigenvector of the mutation-selection matrix (w_{ji}). This eigenvector corresponds to the largest eigenvalue of (w_{ji}), which is the average fitness ϕ [14]. The error threshold is dependent on the existence of a mutation-selection balance when the effect of mutation does not exceed that of the selection pressure; it is the maximal mutation rate that allows a population to stay clustered around the fitness peak. Note that Equation 2.1 is a model for infinite populations. So, strictly speaking, the error threshold does not exist when $N < \infty$. However, Equation 2.1 can be used as an approximation for finite population dynamics [90]. The dynamics of finite populations have been studied for a long time in single-peak landscapes [91, 92]. They have also been studied using the Moran process [93, 14]. The discrete-time formulation of the quasispecies equation has been used to describe mutation-selection dynamics [15, 34, 94].

The concept of an error threshold was initially thought of in terms of genotypes. However, due to neutrality, genotype to phenotype mapping can be many to one. To

maintain a specific phenotype, the system does not need to necessarily maintain one unique genotype. It therefore makes more sense to consider the error threshold in terms of phenotypes [89].

3.1.4.1 Phenotypic Error Threshold

The quasispecies equation can be altered to describe the frequency of phenotypes as opposed to genotypes; this is done by letting one variable represent the group of genotypes that produce the same phenotype, taking into account neutrality (see section 3.1.3) [89]. The phenotypes can be split into two classes; these are the focal phenotype (the wild-type), x, and the class of all the mutant phenotypes, y. As concentration is currently on the maintenance of the focal phenotype in the population, it is only necessary to look at neutral and deleterious mutations. The assumption can be made that any mutation that is not deleterious must be neutral. The effective replication accuracy (Q_e) is given as:

$$Q_e = Q + \Lambda(1 - Q) \tag{3.1}$$

Λ is the fraction of mutants of x that are neutral and therefore do not affect the phenotype. It is also assumed that there is no epistasis, i.e., each mutation will have an independent effect on the phenotype, known as the *additive assumption*. The value of the effective replication accuracy can be calculated based on the length of the sequence and the number of mutations. The minimum per base probability of correct replication (q) for which x can be maintained is given as

$$q_{min} = (\sigma^{-1/L} - \lambda)/(1 - \lambda) \tag{3.2}$$

σ is the superiority of the focal genotype. This is generally calculated as the ratio of the fitness of the highest peak in the landscape to the average fitness of all of the other peaks. In a landscape with one focal genotype (single-peak), this will simply be the replication rate of that focal genotype. λ represents the fraction of neutral mutations out of all possible mutations. L represents the length of the replicating sequences. The error threshold is given as $1 - q_{min}$. Once this has been exceeded, mutation will occur at a high enough rate to cause loss of the focal phenotype. If the per base probability of correct replication (q) decreases, then the overall number of mutations per replication (d) will increase. The higher the number of mutations per replication, the lower the probability of completely neutral replication (λ^d). There is therefore a limit on how much the error threshold can increase for each value of λ (Figure 3.3) [89].

When λ is large ($= \sigma^{-1/L}$), the value of q_{min} becomes equal to 0 [89]. Not only is this value of λ too high to be realistic, the value of q will be so small that there will no longer be a binomial distribution. Takeuchi [89] concluded that this inaccuracy only occurs when $\lambda >0.8$, therefore overestimation of the error threshold using equation 3.2 will only occur when λ is unrealistically high.

The phenotypic error threshold has also been derived by Reidys [2]:

$$q_{min} = \left(\frac{1 - \lambda\sigma}{(1 - \lambda)\sigma} \right)^{1/L} \tag{3.3}$$

Equation 3.3 was derived by Reidys [2] from equation 3.1, and is shown graphically in Figure 3.4. Takeuchi et al. [89] used the formula derived by Reidys [2] to verify that their formulation was correct; both formulations appear to be consistent (refer to Figure 3.3 and Figure 3.4). However, Takeuchi et al. suggest that equation 3.1 can

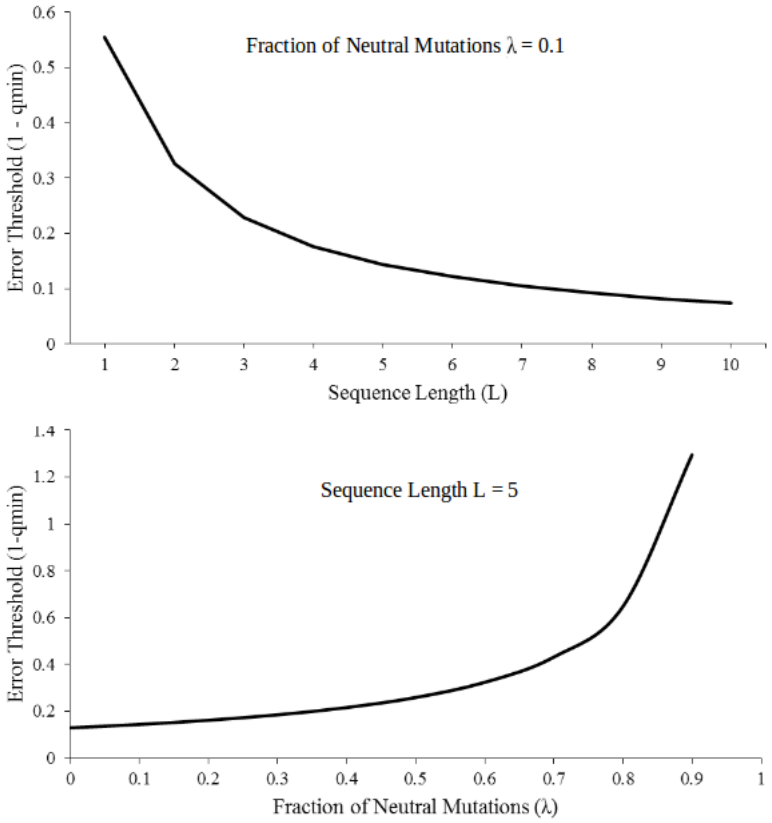

Figure 3.3: **Error threshold $(1 - q_{min})$ for varying values of L and λ, using the equation $q_{min} = (\sigma^{-1/L} - \lambda)/(1 - \lambda)$ [89].** Increasing the sequence length of genotypes will decrease the error threshold, while increasing the fraction of all mutations that are neutral will increase the error threshold.

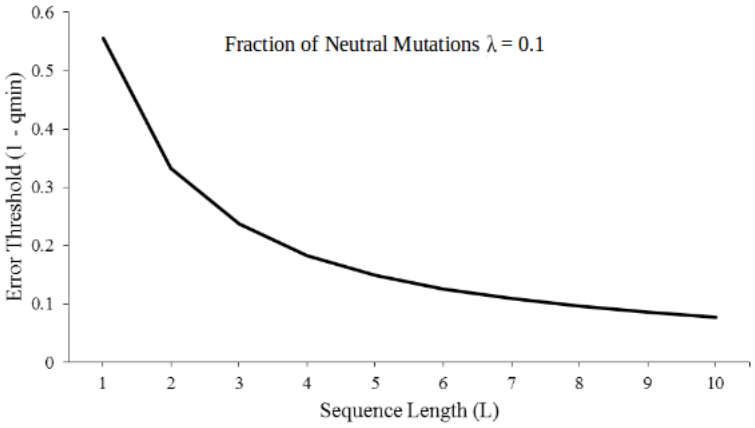

Figure 3.4: **Error threshold $(1 - q_{min})$ for varying values of L, using equation 3.3[2].** Increasing the sequence length of genotypes will decrease the error threshold; this is consistent with the equation for q_{min} that was derived by [89] (refer to equation 3.2 and Figure 3.3).

only be valid if either of the following is true:

- The value of the per base probability of correct replication (q) is large enough so that, for most mutants, the number of mutations per replication (d) is equal to 1.

- There is a neutral set uniformly distributed over the genotype space (where a neutral set is the name given to a set of genotypes which all map to the same phenotype, i.e, its members all belong to the same neutral network).

Takeuchi et al. [89] used equation 3.2 to obtain the following equation for the maximum permissible sequence length (the information threshold):

$$L_{max} = \frac{\ln(\sigma^{-1})}{\ln(q + (1 - q)\lambda)} \tag{3.4}$$

Studies with RNA folding, which have a well studied genotype-phenotype mapping, showed that the average value of λ is a decreasing function of sequence length; the lower the fraction of neutral mutants there are, the greater the maximum permissible sequence length. This relationship means that mutational neutrality imposes a limit on the increase in the value of the information threshold. Takeuchi et al. also backed up their analytical predictions by comparing them with the results of computer simulations [89].

Consider a neutral network with a very large value for q (per base probability of correct replication), in which the majority of mutants have a value of d (overall number of mutations per replication) equal to 1. If the error rate $(1 - q)$ gets close to the error threshold, the average value of d may be greater than 1 for each mutant, even if there are no neutral mutations. The average value of d for each neutral sequence per replication will be lower than that for each replication over all; the error threshold may be overestimated if only one single mutation is considered. To try and improve the estimation of the error threshold, a method called "four λ approximation" can be used instead . This sub-divides each sequence into four, to take into account the fact that the fraction of neutral mutations (λ) may differ depending on the position in the sequence. This reduces overestimation of the error threshold, which is particularly likely to occur when the value of d is large enough to approach that of the error threshold itself [89].

If a mutant sequence is assumed to be neutral (under the additive assumption, which assumes each mutation has an independent effect on the phenotype), but there is actually interaction between its mutations which have a deleterious effect, this is known as *negative epistasis*. The opposite of this is *positive epistasis*. Epistasis can result in mutations that are individually deleterious but jointly beneficial [95]. If there

is a neutral mutant which is a member of the same neutral network as the original sequence (i.e., is only one or two mutations away), this is known as an *additive neutral mutant*. There can also be *additive deleterious mutants*. The additive assumption tends to underestimate the amount of mutational neutrality in a population and the effective replication accuracy (Q_e), although the latter is only noticeable when the error rate exceeds the error threshold. In general, the additive assumption correctly estimates Q_e, and consequently also correctly estimates the error threshold. This estimation remains accurate despite the fact that folding of RNA sequences can lead to failure of the additive assumption, interaction of multiple mutations affecting the phenotype, and therefore epistasis. However, the number of mutations that occur in each neutral replication is sufficiently small so that epistasis is rare. If longer sequences are considered (increased L), the average value of d per neutral replication at the error threshold decreases when λ is constant. As L is increased, the value of λ will decrease. Increasing the sequence length therefore decreases the average d per neutral replication [89].

The error threshold represents the mutation rate above which there is an error catastrophe and the population delocalises across sequence space. Below this, adaptation will occur at a rate dependent on the amount of variation introduced by mutation on which selection can act. Specifically, there is an optimal mutation rate at which adaptation will occur with optimal efficiency, within a limit imposed by the error threshold.

3.1.5 Optimal Mutation Rate and its Relation to Error Threshold

The most sensitive of parameters in a genetic algorithm (GA) is thought to be the mutation rate [96, 97]. Eiben at al. [98] describes two optimization techniques in

GAs: optimal parameter tuning and optimal parameter control (finding a constant optimal parameter value, and starting with an intial parameter value which is varied over time respectively). With respect to a constant value, $1/L$ has been suggested as a general value for the per bit mutation rate in a GA, where L is sequence length [99, 97]. Mühlenbein [99] states that $\mu = 1/L$ is optimal for general unimodal functions (where a function $f(x)$ is unimodal if for a value y, it is monotonically increasing for $x \leq y$, and monotonically decreasing for $x \geq y$) [97]. Ochoa [97] uses GAs, in which the sequences are represented as bit-strings, to test the limitations of the $1/L$ heuristic. The study observed the change in optimal mutation rates with time, along with the interaction between mutation rate, selection pressure, and population size. It was found that a mutation rate of $1/L$ will produce optimal or near optimal results. It was also found that increasing the selection pressure increases the magnitude of optimal mutation rates, with some decrease in optimal mutation rate at small population sizes. It was concluded that a rate of $1/L$ will only be sub-optimal when the selection pressure is either extremely weak or extremely strong, or when the population size is very small [97]. Cervantes and Stephens [100] show that universal heuristics such as $1/L$ and the error threshold can be improved upon when there is information about the underlying fitness landscape. They suggest that both the $1/L$ and error threshold heuristics are too high in landscapes with multiple peaks; the error threshold is dependent on the fitness landscape and has been shown to be related to the optimal mutation rate [80, 100].

Optimal mutation rate is not a precise value, but rather a range of values that allow adaptation to occur with optimal efficiency. This is in contrast with the error threshold, which is thought to be more precise. There is biological evidence that there is a relationship between error threshold and optimal mutation rate. For example, work by Eigen and Schuster [10] showed that viruses live very close to the error threshold;

viruses are known to be very efficient at evolving in new environments [80].

Ochoa et al. [80] empirically tested the hypothesis that optimal mutation rate is related to error threshold. They did this by independently identifying the optimal mutation rate and error threshold *in silico*, and then comparing the values. To validate their empirical method they used existing equations to calculate analytical values for the error threshold, and compared these with values obtained empirically (with the same parameter values). There was only a slight difference in values which could be accounted for by the fact that the analytical method assumed an infinite population, whereas the empirical method was finite [80].

Clune et al. [101] used computer simulations to show that natural selection does not always effectively evolve optimal mutation rates for adaptation in the long-term, and is particularly bad when evolution is occurring on a rugged fitness landscape. On extremely smooth fitness landscapes, mutation rates evolve that are close to optimal. However, mutation rates that are lower than the optimum are favoured when the fitness landscape is rugged with wide valleys; a lower mutation rate is favoured due to the inaccessibility of mutations that are immediately beneficial. Such cases are often made worse by the fact that if there is a low mutation rate, then the chance of a beneficial mutation occurring will be lower than if the mutation rate was high; the process is self-reinforcing in that lower mutation rates lead to fewer beneficial mutations, while reduction in the number of beneficial mutations means any mutation that occurs is more likely to be deleterious and so lower mutation rates are favoured. Experiments with organisms such as yeast, viruses, bacteria and higher eukaryotes have suggested that fitness landscapes are often rugged in the real world, therefore Clune et al.'s results *in silico* are likely to also be relevant in nature. The landscape on which evolution is occurring determines how close to the optimal mutation rate the population will get

[101].

3.1.5.1 Mutation Rate and Selection Pressure

Mutation rate is the most sensitive of the main parameters that affect the performance of a genetic algorithm. The optimal rate per locus is dependent on the reciprocal of the genotype length $(1/L)$ [102, 103]. It has been suggested that optimal mutation rate is proportional to selection pressure; as selection pressure increases, so does the optimal mutation rate [102, 103]. Error threshold is related to optimal mutation rate, and is therefore also dependent on selection pressure. The fitter an organism, the more likely it is to be selected. There needs to be a balance between selection and mutation and recombination; this is the *exploitation-exploration balance*. The strength of a selection mechanism is indicated by the selection pressure, which refers to the change in fitness caused by environmental conditions. It roughly measures the maximum fitness to average fitness ratio within the population. One selection mechanism is rank selection. Each individual in the population is given a ranking based on their relative fitness. Linear ranking puts the individuals in order of fitness, with the lowest at 1, up to the highest at population size N [103].

Imagine a fitness landscape with a single peak with fitness $\sigma > 1$, and with all other sequences having fitness 1. The error threshold, denoted by Ochoa et al. [103] as p, is given as:

$$p = \frac{\ln(\sigma)}{L}$$

σ is the selective advantage of the master sequence over all other sequences in the population, i.e., the selection pressure. L is the sequence length. In the simplest

instance, σ represents the ratio of the reproduction rate of the master sequence (its fitness) to the average reproduction rate of the rest of the sequences in the population. Ochoa et al. [103] looked at error thresholds and optimal mutation rates when solving both toy and real world problems with a GA. They concluded that error thresholds and optimal mutation rates are correlated. In addition to this, the strength of the selection pressure was shown to have a significant effect on the magnitude of both the error threshold and the optimal mutation rate; the stronger the selection, the greater the values of error threshold and optimal mutation rate [103].

3.1.6 Mutational Robustness and Survival-of-the-flattest

In addition to the error threshold, in landscapes where there is more than one peak, there may also be one or more critical mutation rates at which the population loses its ability to remain on fitter peaks, but retains its ability to remain on flatter peaks of lower fitness [33, 26, 1, 34]. Above such critical mutation rates, individuals with greater robustness to mutation are able to survive while fitter, less robust individuals may not. This represents a phase transition from survival-of-the-fittest to survival-of-the-flattest [33, 13, 1, 34, 21, 35, 29]. At lower mutation rates, selection favours individuals that reside at peaks with higher fitness, due to the rarity of mutations that push individuals off the peaks [20]. However, at higher mutation rates there will be an increase in the frequency of mutations which push individuals off the peaks; selection favours individuals located in flatter regions of the fitness landscape as individuals here are less likely to experience large reductions in fitness compared with those that may be initially fitter but reside in parts of the landscape with steeper peaks. Individuals that are part of a neutral network [104, 105, 106] (see section 3.1.3), in that they are connected in sequence space to other individuals with equivalent fitness, are said to be

more mutationally robust than individuals that are not [13, 107, 108, 37]. The critical mutation rate has been defined as the midpoint between the highest mutation rate at which there is survival-of-the-fittest, and the lowest mutation rate at which there is survival-of-the-flattest [33, 1].

Survival-of-the-flattest has been observed in digital organisms [33, 35], theoretically [108, 35], in simulated RNA evolution [37], and in RNA viruses [21, 109]. In addition, evolution of mutational robustness has been observed in simulated RNA evolution [110], and in laboratory protein evolution experiments [111]. Both van Nimwegan et al. [110] and Bloom et al. [111] place an emphasis on the degree of polymorphism in the population, suggesting that highly polymorphic populations are more likely to spread across many nodes of a neutral network (each corresponding to a genotype), concentrating at highly connected parts; the members of the population at highly connected nodes have greater robustness to mutation, which they pass on to the next generation. Robustness will evolve in any population where the product of the population size and mutation rate is sufficiently large (>1). Krakauer and Plotkin [36] refer to robust landscapes as redundant, and less robust, steeper landscapes as antiredundant. They suggest that both in theory and in individual-based stochastic simulations, redundancy increases the mean fitness in small populations as it masks the mutations that arise due to mutational drift. However, large populations are less affected by drift, and so it is advantageous to have a sharp landscape with a high maximum fitness.

3.2 Small Populations and the Risk of Extinction

Small populations frequently exist in nature. Some animal species can exist in populations of only hundreds, while those nearing extinction may be found in populations of only a few individuals. For example, the Chatham Island Black Robin was recorded

as existing in two populations consisting of 190 and 34 mature individuals as of spring 2011 [4], only four of the 15 known populations of cheetahs in Eastern Africa were estimated to consist of greater than 200 individuals as of 2008 [5], and Campbell's Alligator Lizard has been reported as having a total estimated population size of 500 individuals as of 2010 [6]. Populations at risk of extinction are of particular concern. Understanding the effect of population size on the critical parameters of evolution (mutation, recombination, selection, and genetic drift) is essential in making accurate predictions regarding the likely fate of a small population if left to persist in its current environment. For example, inbreeding resulting from genetic drift in small populations can depress population fitness and increase the risk of extinction [112]. Environmental change is rapid, therefore populations need to evolve at a sufficient rate to prevent further population decline and enable evolutionary rescue [7]. Population decline can lead to loss of fit genetic material that may be difficult to recover in very small populations due to mutational meltdown [8]. Meltdown occurs when a deleterious mutation becomes fixed in a population leading to reduced fitness and therefore reduction in population size. Mutations become fixed more rapidly the fewer individuals there are in the population; each time fixation of a deleterious mutation leads to reduction in population size it becomes easier for further deleterious mutations to become fixed leading to a potential downward spiral towards extinction.

3.2.1 Extinction Thresholds

The error threshold does not necessarily equate to an extinction threshold [113]. An error catastrophe is an evolutionary shift in genotype space, while extinction refers to the reduction of individuals in the population. A population that shifts to the lower fitness areas of the landscape is less well adapted to its current environment. However,

if the population shifts to genotypes that are robust to mutation, this can delay or prevent extinction due to survival-of-the-flattest [113].

Extinction is caused by the process of lethal mutagenesis, through mutational meltdown and the process of Muller's ratchet [113]. In a finite population with no back mutation, Muller's ratchet is the process by which deleterious mutations accumulate until every individual contains at least one deleterious mutation, resulting in complete loss of the wild-type. If this process continues, each individual accumulates an increasing number of mutations so that progressively more sequences with fewer mutations are lost from the population. In lethal mutagenesis, mutation reaches a high enough rate so that there are too many deleterious mutations for the population to maintain itself [113]. This is subtly different from an error catastrophe during which higher fitness genotypes are lost as they are too sensitive to the rate of mutation, with less fit but robust genotypes taking precedence. In the context of viruses, Bull et al. [113] present a condition for lethal mutagenesis in the absence of an error catastrophe:

$$e^{-U_d} R_{max} < 1$$

U_d represents mutation rate, while e^{-U_d} is both the mean fitness level and also the fraction of offspring with no non-neutral mutations. R is the average number of viable offspring per cell infected by the wild-type genotype, while R_{max} denotes the maximum reproductive rate of the wild-type genotype. This equation represents an extinction threshold [113].

This chapter has covered the different types of mutation, the nature of adaptation, optimal mutation rates and error thresholds, and the concept of mutational robustness and survival-of-the-flattest. Finally, factors associated with small populations and the risk of extinction were discussed. This review of the current state of the literature

has identified some questions that require answers. These are briefly introduced in chapter 4 along with a discussion of the methodology and model design required to answer them. Each question is discussed in detail, along with the reasoning behind it, in subsequent chapters.

Chapter 4

Research Questions and Methodology

4.1 Research Questions

Survival-of-the-flattest has been observed in digital organisms [33, 35], theoretically [108, 35], in simulated RNA evolution [37], and in RNA viruses [21, 109]. Evolution of mutational robustness has also been observed in simulated RNA evolution [110], in an artificial evolution model with digital organisms [114], and in laboratory protein evolution experiments [111]. Both Wilke [37] and Comas et al. [1] found "that population size played only a minor role in determining the position of the critical mutation rate" [34], within the context of their experiments. Population sizes as low as 250 were used, and the conclusion made "that the critical mutation rate was independent of population size" despite the fact that there did appear to be some correlation for certain cases [1]. They did not consider smaller populations, such as those that may exist for species nearing extinction or living in localized groups. Both Nowak and Schuster [32] and Wiehe et al. [38] considered the effect of random genetic drift in finite populations (in haploids and diploids respectively), and observed that there is a shift of error thresholds to lower values which is more pronounced the smaller the population. Error

thresholds were also shown to increase for increasing population size using a genetic algorithm with both single-peak and correlated landscapes [40]. These results, along with a review of the current state of the literature, pose the following questions:

1. Does the critical mutation rate ever vary with population size in a haploid genetic algorithm?

2. Does the critical mutation rate ever vary with population size in a diploid population replicating using a system modelled on the biological process of meiosis?

3. Does the critical mutation rate vary in a haploid system compared with a diploid system, and if so, is this due to the difference in recombination?

4. Is the output of the artificial simulation model relevant to a real biological population?

Chapter 5 discusses question 1, chapter 6 discusses questions 2 and 3, and chapter 7 discusses question 4. The reasoning behind each question is given along with hypotheses, experiments, results and conclusions.

4.2 Simulation Models

To answer the questions above, a system was required to model the simplest case in which survival-of-the-flattest could occur, in which the results produced were independent of the details of the underlying landscape, in which parameter values could be easily manipulated, and which could run a simulation within a reasonable time frame. Here two existing artificial life systems are examined along with their limitations with respect to the aims of this work, and the design of the new simulation model is de-

scribed. The simulation model is subsequently developed to more closely resemble a biological system with a diploid population.

4.2.1 Existing Artificial Life Systems

Evolution experiments can be carried out in a range of systems from biochemical organisms to standard Monte-Carlo simulations [115]. Systems with self-replicating entities provide a half-way point between the two, allowing controlled (potentially open-ended) evolution experiments to be done *in silico*. Two existing systems that evolve digital organisms are Tierra and Avida.

4.2.1.1 Digital Evolution *in silico*

The Tierra software (http://life.ou.edu/tierra/index.html) [116] creates a virtual computer in which the executable machine codes are evolvable. The machine code can be mutated by flipping bits and recombined by swapping segments of code between algorithms, with resulting code subject to a form of natural selection. The operating system provides control of factors including mutation rate, disturbances, and soup size. It also keeps track of births and deaths, sequences, successful genomes, and interactions between creatures.

The Avida software (http://avida.devosoft.org/) was inspired by the Tierra system. Avida has an update mechanism similar to a 2D cellular automata [117]. Like Tierra, the creatures are evolvable strings of machine code. Each creature has an independent co-ordinate position in a "grid" that marks its physical location, where the points of the grid can interact with each other in a similar manner to a cellular automata. Unlike a cellular automata, the update rules are not fixed but rather depend on the genomes themselves. The genomes self-replicate, determine their own size, and allocate memory

accordingly [117]. They can mutate and, like Tierra, birth and deaths are recorded. Both the Tierra and Avida systems divide genotypes into *threshold* and *temporary* genotypes. Genotypes in Avida are classed as threshold if more than 10 members of the population have that genotype concurrently, while all other genotypes are classed as temporary. The threshold is higher in Tierra and consequently Tierra does not record short-lived genotypes. Extinction events are less severe than those in Tierra [117]. Avida has been used in evolution studies focusing on mutation rates, e.g., [33, 101].

Such artificial life systems behave as an evolving system, and have parallels to organic evolution on earth [118]. According to Ray [118]: "If we accept that evolution is the defining property and creative process of life, then instances of digital evolution may also be considered instances of life, albeit dramatically alien life". He describes how artificial life in a digital medium shares only the actual process of evolution with life on Earth, the advantage of this being that digital evolution gives a broader perspective on what evolution is and what it does, independent from details of the chemical and biological processes of the mechanisms. One of the aims of this work is to produce results that are of potential relevance to a real biological system. This includes consideration of the underlying mechanisms such as haploidy and diploidy and the use of parameter values dependent on the interactions between biological molecules, e.g., mutation rates observed in nature. It therefore would be beneficial to design a system from the bottom up; the system should be based on the mechanisms present in biology that, when run, evolves, as opposed to an evolving system in which the exact mechanisms of evolution are unimportant.

4.2.2 A Simple System with Survival-of-the-Flattest

Jones and Soule [19] determined that the role of genetic robustness in evolution differs significantly depending on whether it is a generational or steady state genetic algorithm that is being used. Studies have confirmed the notion of survival-of-the-flattest using generational models, such as Wilke et al.'s [33] evolution of digital organisms in Avida, and Krakauer and Plotkin [36] study of redundancy and antiredundancy [19]. Jones and Soule [19] suggest that for evolutionary dynamics experiments, the class of algorithm used can have a significant effect on the observed outcome. They point to steady state algorithms as being of particular interest to the artificial life community, as evolution in microbes resembles the action of a steady state-like algorithm. However, the problem with steady state algorithms is that they can allow individuals to survive on fitness peaks indefinitely. This is not a realistic property when modelling evolutionary dynamics; a preferable approach is to use a generational genetic algorithm which retains the key features of steady state evolution: selection in which individuals with a higher fitness score beat those with a lower fitness score, and in which some degree of asynchronicity is retained. It should be noted that fitness in this sense refers to a score assigned to each individual based on a given fitness function, as opposed to the biological definition of fitness as a measure of replication rate; the exact fitness values used are unimportant as it is relative fitness that determines which individuals are selected. Specifically, selection is determined based on whether the fitness of one individual is negative or positive relative to another, regardless of the size of the fitness difference. This approach also allows for the existence of a critical mutation rate: with a standard steady state algorithm, always retaining the fittest individual prevents the population from ever losing the highest current peak.

The following section describes the Haploid Method designed and used to ascertain

whether the critical mutation rate ever varies with population size. This is followed by a modified form of the simulation model in which an algorithm is developed that follows the biological process of meiosis with diploid organisms. This is devloped further to improve efficiency. Development of these methods was done through design and implementation of a genetic algorithm [39]. The Diploid Method was used to answer research questions 2 and 3, while the Diploid Method with Improved Efficiency was used to answer research question 4. The source code used will be available electronically.

4.2.2.1 Haploid Method

An individual sequence consists of a string of characters drawn from an alphabet of size 4 (which can be thought of as, for example, A/C/G/T or 0/1/2/3) with a fixed length of 30. In each step of the algorithm, three individual sequences are selected at random from the population. Two of the three selected individuals are chosen as parents in a crossover which replaces the third individual with the resulting child. The child is then subject to one round of point mutation (to a *different* base) at a given per-base mutation rate. The individual to be replaced is determined each time based on the fitnesses of the three selected individuals: there is an equally small chance of either of the two fittest of the three being replaced (25%), and a larger chance of replacing the least fit (50%) The 25:25:50 ratio ensures that any individual can be chosen, so allowing a population to lose its fittest peak. This use of tournament selection ensures that selection is independent of the precise shape of the landscape. This process continues until each individual in the population has been chosen exactly once (or there are less than three remaining to select); this represents one generation. The fitness of each individual sequence is evaluated based on a two-peak fitness landscape with one narrow peak of high fitness (peak 0), and a broader, flatter peak with lower fitness (peak 1)

(Figure 4.1). Peak 0 has a maximum fitness score of 15 and a radius of 2, where radius refers to the Hamming distance from top-of-peak to zero fitness score. Peak 1 has a maximum fitness score of 10 and a radius of 5, with its top chosen as an arbitrary point (fixed throughout evolution) with a Hamming distance of 10 from the top of peak 0. This is done by setting the sequence at the top of peak 0 to be a string of 0s, while the sequence at the top of peak 1 is set as a string of 0s with 10 of those 0s randomly changed to either a 1, 2 or 3. Individuals are allowed to move on the slopes, or in between the peaks. This is a simple landscape in which survival-of-the-flattest can occur, with generality due to the use of tournament selection. The effect of mutation on fitness is smaller within peak 1 than within peak 0; individuals located on peak 1 will have higher mutational robustness compared with those located on peak 0.

Following the experimental procedure designed by Wilke et al. [33] (and used by Comas et al. [1]) half of the population was initialized to peak 0 and half to peak 1 to avoid any initial bias towards either peak. The simulation was run for 10,000 generations, and the first generation at which there were no individuals on peak 0 was recorded. At this point, all the individuals were 2 or more mutations away from the top of peak 0 and therefore peak 0 was considered to have been lost (Figure 4.1). If peak loss did not occur within the 10,000 generations, a value of -1 was recorded in place of the generation number. Similarly, the number of generations it took to lose peak 1 was also recorded, where peak 1 was considered to have been lost when all individuals were 5 or more mutations away from the top of peak 1 (Figure 4.1). A range of per-base mutation rates was tested for a range of population sizes, with the simulation being repeated and run 2000 times for each combination. It should be noted that the population size is fixed for the duration of each run. The mutation rate by which 95% of the runs had lost each peak was recorded as a critical mutation rate,

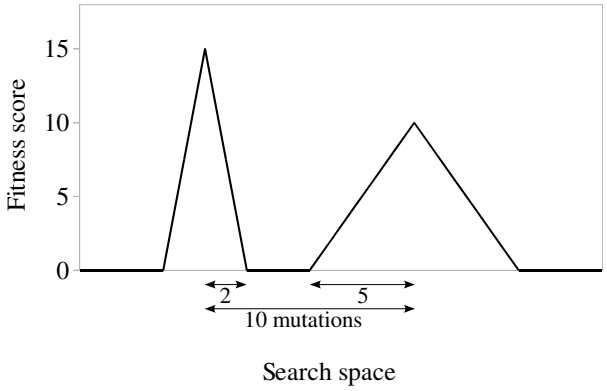

Figure 4.1: **Two-peak fitness landscape**, with one narrow peak of high fitness (Peak 0), and one broader peak of lower fitness (Peak 1). Diagram adapted from Wilke [34].

where a peak was considered to have not ever been lost only if there were individuals remaining on it at the end of the 10,000 generations. This number of generations was chosen after preliminary runs determined this to be long enough for the population to localize to either peak or disperse, yet feasible in terms of runtime.

4.2.2.2 The Moran Process

Evolution in finite populations can be described using the Moran process [14]. The Moran process is a simple birth-death process in which each time step involves choosing at random one individual for reproduction and one individual for death. Death occurs by replacing the latter individual with the child of the former individual. There are no restrictions to ensure each individual is chosen any number of times. Nowak and Schuster use a system based on the Moran process [32], in which they group individuals into error classes (where individuals are in the same error class if they are the same Hamming distance away from the master or target sequence). They take into account the number of individuals in each error class to calculate the transition probabilities

of the birth and death process. The haploid method described above is a variation on this; population mixing is done through crossover of two reproducing sequences to replace a third sequence marked for death, with every individual being chosen once and only once to provide a defined generation. Use of crossover to introduce mixing is a more biologically realistic process. In addition, while Nowak and Schuster's method considers frequency at the level of the population, the haploid method described above operates at the level of the individual sequence.

4.2.2.3 Diploid Method

Modelling not only evolution, but the process by which evolution occurs in nature, is a step towards bridging the gap between artificial and biological evolution. The genetic algorithm for a diploid population was modelled on the biological process of meiosis. Meiosis is a type of cell division which produces haploid cells. DNA is made up of two complimentary sequences (double-stranded), and condenses during cell division to form structures known as chromosomes [51]. Diploid organisms have two copies of each chromosome. For example, humans have 46 chromosomes in 23 pairs. One of each pair comes from the father, the other from the mother. The pairs are known as homologues. Each chromosome is replicated (during which process there is a chance of mutation), and subsequently becomes a complex made up of two identical sister chromatids which form an X-shaped structure (bivalent). Homologous chromosomes join together to form a tetrad. This means, for example, that the maternal copy of chromosome 1 will pair up with the paternal copy of chromosome 1. It is called a tetrad as it is made up of four chromatids (the original maternal and paternal chromosomes and their duplicates). Crossover occurs within the tetrads. The pairs are pulled apart to opposite ends of the cell. Each end of the cell will subsequently have one copy of

the chromosome. The cell splits to create two cells, each with the correct number of chromosomes (one copy of each). Each cell will contain a mix of paternal and maternal DNA due to crossover. In each of the two cells, the chromosomes are split into their constituent chromatids. The chromatids are pulled to opposite ends, and the cells divide. The result is four cells, each containing one chromatid (now referred to as a chromosome). The resulting four cells are haploid as they contain only one copy of each chromosome, and are known as gametes. The joining of a gamete from a mother with that from a father will produce a diploid child.

In the genetic algorithm, each genetic sequence is represented as a string of 30 characters. DNA is double-stranded, but as one strand is just a compliment of the other, it can be represented as a single-strand string in the simulation. Consistent with the haploid system, each character in the sequence is one of four possibilities. A diploid individual consists of two sequences, one inherited paternally and one inherited maternally. There are no distinct sexes in the simulation; the terms *maternal* and *paternal* are used merely to differentiate between the two parent individuals. At the start of the algorithm, each peak is initialized so that it contains half of all of the sequences. Overall, a quarter of the population has both its maternal and paternal sequences on peak 0, a quarter have both their maternal and paternal sequences on peak 1, and the remaining individuals consist of maternal and paternal sequences on different peaks. This was done to ensure that half of the individuals on each peak were homozygous while the other half were heterozygous. In each step of the algorithm, three individuals are selected from the population at random. Two of the selected individuals are chosen to be parents, while the third will be replaced by their child after reproduction. Selection is carried out based on the fitness of the three individuals. There is an equally small chance of either of the two fittest individuals being chosen

to be replaced (25%) and a higher chance of the individual with the lowest fitness being replaced (50%). After selection, crossover occurs within each parent individual between the maternal and paternal sequences. A locus is randomly selected to be the crossover point. The maternal sequence is copied up to this locus, and the paternal sequence after. This produces a single-sequence gamete from each parent, the bases of which are then mutated (each to a different base) according to a per-base mutation rate. One of the gametes is randomly designated the paternal sequence for the child, while the other becomes the maternal. The resulting diploid child becomes part of the population. This process continues until each individual in the population has been chosen exactly once (or there are less than three remaining to select); this represents one generation and ensures that there is no chance of any individual avoiding being chosen and so remaining static in the landscape. The fitness of each individual sequence is evaluated based on the two-peak fitness landscape (Figure 4.1) and the experimental procedure is that used in the haploid system.

4.2.2.4 Fitness Calculation

The key difference between the system described in section 4.2.2.1 (and in [41]) and the system described in section 4.2.2.3 is the introduction of diploidy. In the haploid case, the fitness of each individual is calculated based on the Hamming distance of an individual sequence from the top of each peak. The fitness of the individual in terms of peak 0 is equal to $\max(0, f_0 \times (1 - d_0/r_0))$, where f_0 is the fitness score of the target at the top of peak 0, d_0 is the Hamming distance of the individual from this target, and r_0 is the Hamming distance between the target and the point at which the peak has a fitness score of 0 (see Figure 4.1). The fitness of the individual is also calculated in terms of peak 1. The higher of the two fitness values is designated to be the overall

individual fitness score. However, a diploid individual consists of two sequences and has a fitness score for each. To obtain an overall individual fitness score, f, a dominance parameter, λ, is introduced:

$$f = (\lambda \times f_{\max}) + ((1 - \lambda) \times f_{\min}) \tag{4.1}$$

The fitness score for each of the constituent sequences is compared. The sequence with the higher of the two fitnesses has its fitness score designated $f_{\max}$, while $f_{\min}$ is the fitness score for the sequence with the lower fitness. If both sequences have the same fitness, $f_{\max}$ and $f_{\min}$ will have equal value. When λ is set equal to 1, the overall individual fitness is equal to the maximum of the two fitness scores. When λ is set equal to 0, the overall individual fitness will be equal to the minimum of the two fitness scores. The experiment was run with λ set at a range of values, $0 \leqslant \lambda \leqslant 1$.

4.2.2.5 Diploid Method with Improved Efficiency

The diploid method with improved efficiency follows the same algorithm as the diploid method described in section 4.2.2.3 but with adjustments to take away the scaling of runtime with sequence length. The population is no longer stored in a 2D array of dimension population size by sequence length, but rather as an array of vectors, each of which represents a single individual. Instead of storing each individual as a string of characters, the vector contains a list of all the positions in the sequence stored in the form of a structure made up of two values, the position and either a 1, 2 or 3. Positions in the sequence where there is a 0 will not be stored. This means that entire sequences are no longer stored which is an advantage when the sequences are very long. In addition to this, the mutation step has been modified so that mutation is no longer

done base by base and therefore runtime no longer increases with sequence length. Mutation now occurs by generating a random number K from a bionomial distribution of L trials with M probability of mutation, where K represents the number of bases in the sequence that will mutate according to the current rate of mutation. The rest of the algorithm follows the same steps as that described in section 4.2.2.3, with fitness calculation as described in section 4.2.2.4. This method was run with sequence lengths in the range found in biology, as reported in chapter 7.

Chapter 5

Critical Mutation Rates in Haploid Populations

Population size can range from small numbers of individuals to very large numbers of individuals. For example, RNA viruses can reach population sizes of around 10^{10} individuals in a short amount of time [1], whereas some animal species may exist in populations consisting of only hundreds or even tens. The genome of each individual constantly evolves through the processes of mutation, recombination (in sexual reproduction), selection, and genetic drift [54]. Population dynamics can be modelled *in silico* using genetic algorithms, in which populations of sequences are allowed to undergo selection, crossover, and mutation at specified rates; studies can be done in a controlled environment within time-frames not possible in many natural biological systems, producing results that are comparable both to theory and to experimental results in microorganisms.

5.1 Background and Hypothesis

In any evolutionary system, including genetic algorithms and natural biological systems, there is significant evolutionary pressure to evolve sequences that are both fit and robust [19]: in environments with high levels of mutation, robustness can reduce the negative effects of deleterious mutation. Smaller populations are susceptible to random genetic drift [1, 54], therefore it is expected that population size should influence the size of mutation rate that can be tolerated before fitter individuals are outcompeted by those with a greater mutational robustness. Comas et al. [1] used digital organisms to determine whether population size has a direct effect on the position of the critical mutation rate at which the population loses its ability to localise to the fittest peak but retains its ability to remain on lower peaks that have greater mutational robustness. Using population sizes in the same range as those used by Wilke et al. [33], they concluded that there was a lack of general correlation between the critical mutation rate and population size, despite the fact that there did appear to be some correlation for certain individuals in the population.

Survival-of-the-flattest has been observed in digital organisms [33, 35], theoretically [108, 35], in simulated RNA evolution [37], and in RNA viruses [21, 109]. Evolution of mutational robustness has also been observed in simulated RNA evolution [110], in an artificial evolution model with digital organisms [114], and in laboratory protein evolution experiments [111]. Both [110] and [111] place an emphasis on the degree of polymorphism in the population, suggesting that highly polymorphic populations are more likely to spread across many nodes of a neutral network (each corresponding to a genotype), concentrating at highly connected parts; individuals at highly connected nodes have greater robustness to mutation, which they pass on to the next generation. Flat landscapes have been referred to as redundant, and steeper landscapes as antire-

dundant. It has been suggested that both in theory and in individual-based stochastic simulations, redundancy increases the mean fitness in small populations as it masks mutations that arise due to mutational drift [36]. However, large populations are less affected by drift, and so are more able to occupy high-fitness peaks in sharp landscapes.

Both Wilke [37] and Comas et al. [1] found "that population size played only a minor role in determining the position of the critical mutation rate" [34], within the context of their experiments. Population sizes as low as 250 were used, and the conclusion made "that the critical mutation rate was independent of population size" despite the fact that there did appear to be some correlation for certain cases [1]. They did not consider smaller populations, such as those that may exist for species nearing extinction or living in localized groups. Both Nowak and Schuster [32] and Wiehe [38] considered the effect of random genetic drift in finite populations (in haploids and diploids respectively), and observed that there is a shift of error thresholds to lower values which is more pronounced the smaller the population. Error thresholds were also shown to increase for increasing population size using a genetic algorithm with both single-peak and correlated landscapes [40]. Based on these results for error thresholds, the need for further investigation of the critical mutation rate at smaller population sizes than those previously studied is considered, and the following hypothesis is posed:

Hypothesis - Critical mutation rate has a dependence on population size in haploid populations.

5.2 Methods

The hypothesis was tested using the haploid method described in section 4.2.2.1. Following the experimental procedure designed by Wilke et al. [33] (and used by Comas et al. [1]) half of the population was initialized to peak 0 and half to peak 1 to avoid

any initial bias towards either peak. The simulation was run for 10,000 generations, and the first generation at which there were no individuals on peak 0 was recorded. At this point, all the individuals were 2 or more mutations away from the top of peak 0 and therefore peak 0 was considered to have been lost (Figure 4.1). If peak loss did not occur within the 10,000 generations, a value of -1 was recorded in place of the generation number. Similarly, the number of generations it took to lose peak 1 was also recorded, where peak 1 was considered to have been lost when all individuals were 5 or more mutations away from the top of peak 1 (Figure 4.1). A range of per-base mutation rates was tested for a range of population sizes, with the simulation being repeated and run 2000 times for each combination. It should be noted that the population size is fixed for the duration of each run. The mutation rate by which 95% of the runs had lost each peak was recorded as a critical mutation rate, where a peak was considered to have not ever been lost only if there were individuals remaining on it at the end of the 10,000 generations. This number of generations was chosen after preliminary runs determined this to be long enough for the population to localize to either peak or disperse, yet feasible in terms of runtime.

5.3 Results

The results indicate that population size affects the size of mutation rate required for the predominant outcome of the runs to shift from survival-of-the-fittest to survival-of-the-flattest, and that this is particularly noticeable in populations with 100 individuals or fewer. Similarly, the size of mutation rate required for approximately 95% of the runs to have lost both peaks also has a dependence on population size. The results of the simulation can be approximated by a simple exponential function: $y = A - B * x^C$ for some values of the parameters A, B and C. However, they are more closely fitted

by a stretched exponential function: $y = A - B * e^{-((N/C)^D)}$, where N is population size (Figure 5.1).

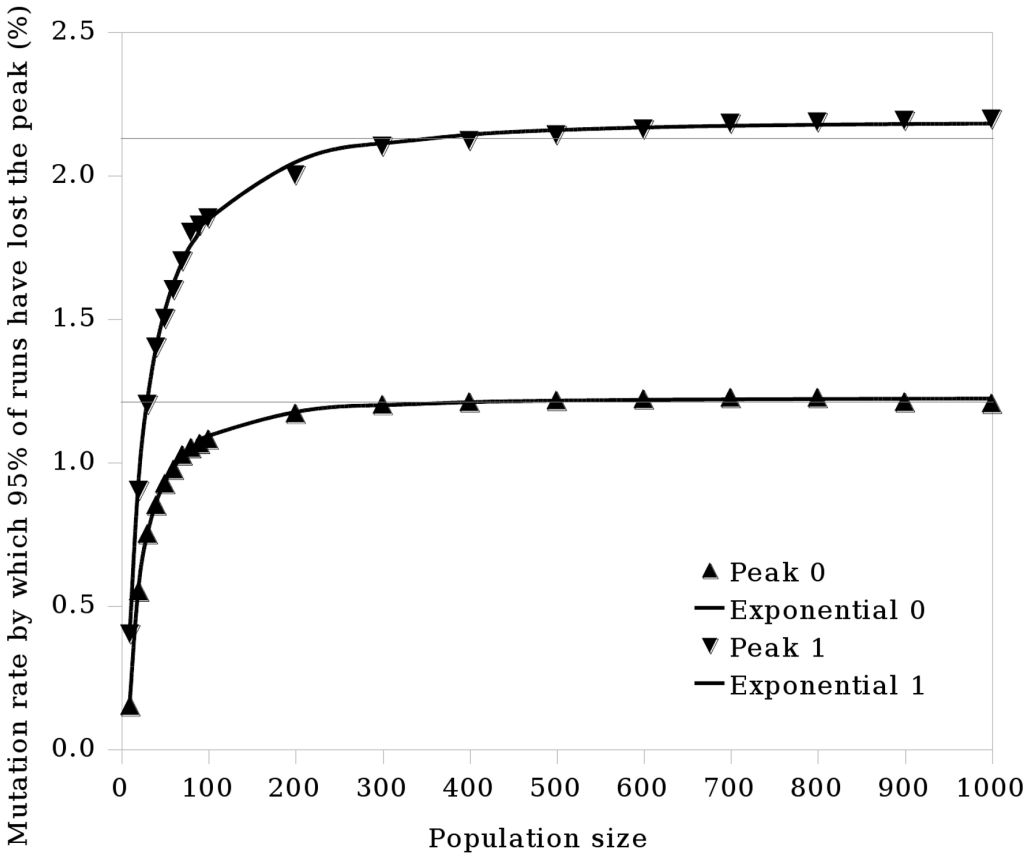

Figure 5.1: **The results of the simulation can be approximated by an exponential function.** This applies to both peak 0 (high, narrow peak) and peak 1 (lower, flatter peak). $y = A - B \times \exp -((N/C)^D)$(with N being population size). The parameters (and their standard error in brackets) obtained by curve-fitting using R with a least squares method were, for the high, narrow peak (peak 0): A = 1.221% (0.0033%), B = 7.001% (1.4390%), C = 1.440 (0.1701), D = 0.3250 (0.0274), and for the lower, flatter peak (peak 1): A = 2.184% (0.0122%), B = 5.438% (1.0466%), C = 7.721 (0.2734), D = 0.3978 (0.0476).

5.3.1 Observed Error Thresholds are Consistent with Analytical Models

An algorithmic method was developed that simulates evolution of a haploid population on a two-peak landscape (see section 4.2 and Figure 4.1). Using this, both the critical mutation rate and the error threshold were measured for a range of population sizes. Nowak and Schuster [32] use a system based on the Moran process and present an analytical expression for the population size dependence of the error threshold (Equation 5.1), where q_{min} is the error threshold, v is sequence length, σ is the selection strength or superiority parameter of the master (fittest) sequence, $\alpha = \sqrt{\sigma - 1}$, and N is population size:

$$[q_{min}(N)]^v = \frac{1}{\sigma}\left[1 + \frac{2\alpha^2}{N}\left(1 + \frac{\sqrt{N}}{\alpha}\sqrt{1 + \frac{\alpha^2}{N}}\right)\right] \tag{5.1}$$

Ochoa et al. [119, 80] derived a reformulation of the Nowak and Schuster analytical expression (Equation 5.2), in which they make explicit the reduction in the error threshold when moving from infinite populations to those of size N (see [119] section 3 for the detailed derivation). Here p_N is the error rate:

$$p_N = \frac{\ln(\sigma)}{v} - \frac{2\sqrt{\sigma - 1}}{v\sqrt{N}} + \frac{2\ln(\sigma)\sqrt{\sigma - 1}}{v^2\sqrt{N}} \tag{5.2}$$

Figure 5.2 shows the error thresholds from our algorithmic method alongside those from Equations 5.1 and 5.2 using a σ value of 2.1. It should be noted that σ is the superiority parameter which would normally be calculated as the ratio of the two fitness peaks. However, as fitness in our algorithmic method is represented as a score as opposed to being a direct measure of reproductive rate, and selection is determined only by fitness score rank, independent of the magnitude of fitness score difference (such that any strictly monotonic transformation of fitness score would produce the

same results), we show here the curves with the σ value that best fits the complete range of our results. The values of 15 and 10 as the peak heights were initially chosen arbitrarily to ensure one peak was higher than the other and the validation with the predictions of Nowak and Schuster were done subsequently. It has been confirmed that changing the original algorithmic method to include peak heights with a ratio of 2.1 produces a comparable match (Figure 5.3). The observed consistency with the analytical models provides verification for the algorithmic method, and therefore confidence in the subsequent results.

5.3.2 Transition from Survival-of-the-fittest to Survival-of-the-flattest

As opposed to there being instantaneous transitions from survival-of-the-fittest to survival-of-the-flattest and to the error catastrophe at discrete mutation rates, there are gradual transitions in which there are shifts from the first to the second, and from the second to the third (Figure 5.4). The mutation rate at which 95% of the runs have lost the high, narrow peak (peak 0) within 10,000 generations marks a point at which the transition from survival-of-the-fittest to survival-of-the-flattest is essentially complete. This can be considered as a critical mutation rate. For a haploid population of 100 individuals of length 30, this is at a per-base mutation rate of approximately 1.08%. Figure 5.4(a) shows the number of generations taken to lose each peak at this mutation rate, for each of the 2,000 runs. Just 52% of these runs lost peak 1 within the duration of the simulation (compared to 95% for peak 0). Loss of peak 0 is then followed by one of two events: either peak 1 is lost relatively quickly (within 200 generations) or it is maintained for the duration of the simulation. The fate of the population after loss of peak 0 is therefore dependent on whether or not it is able to quickly converge on peak

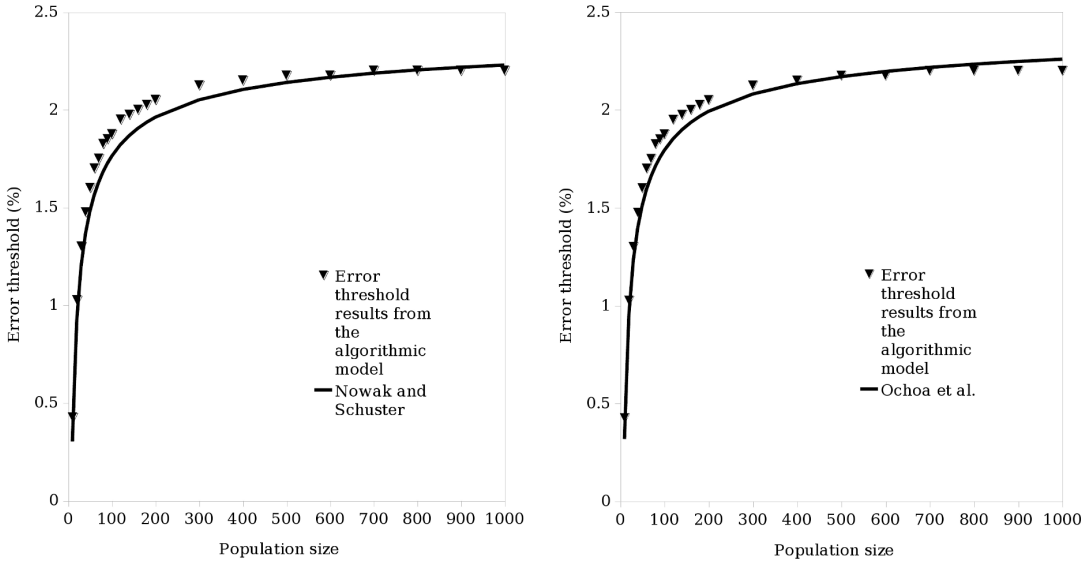

Figure 5.2: **Verification of the method against analytical models for the error theshold.** An analytical expression for the population size dependence of the error threshold is presented (Equation 5.1)[32]. Ochoa et al. [119, 80] include a reformulation of the Nowak and Schuster analytical expression (Equation 5.2), in which they make explicit the reduction in the critical mutation rate when moving from infinite populations to those of size N (see [119] section 3 for the detailed derivation). The observed consistency between our results and the analytical models provides verification for our results and the algorithmic method as a whole. It should be noted that the x axis represents the mutation rate by which 95% of runs have lost the lower, flatter peak (peak 1).

1. Figure 5.4(a) shows (at this mutation rate) that when peak 0 is not lost early, the number of generations taken to lose peak 0 is distributed approximately evenly up to 10,000 generations. The mutation rate corresponding to 95% of the runs having lost the lower, flatter peak (peak 1) within 10,000 generations marks a point at which the transition from survival-of-the-flattest to the error catastrophe is essentially complete. This can be considered as another critical mutation rate (or the error threshold). For a

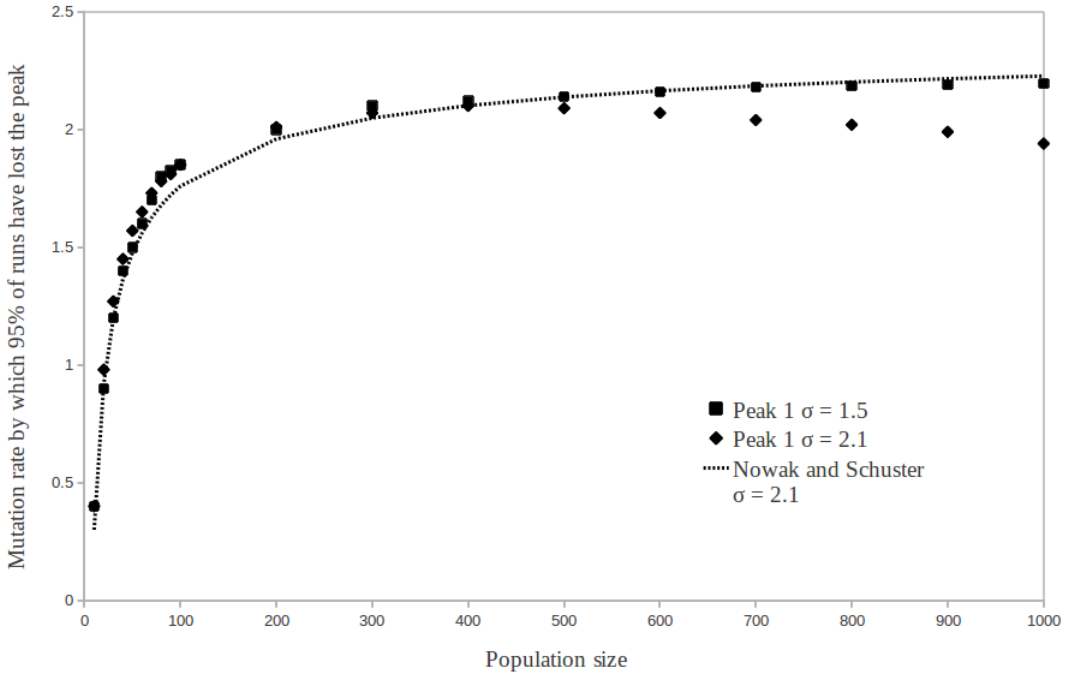

Figure 5.3: **Relevance of superiority parameter σ when fitness is a relative score.** The original results of the algorithmic method (Peak 1 σ = 1.5) and Nowak and Schuster's analytical model (Nowak and Schuster σ = 2.1) as presented in Figure 5.2, but with additional results from the algorithmic method where peak heights had been set to give a σ value of 2.1 (Peak 1 σ = 2.1). It is apparent from this that when σ = 2.1 using Nowak and Schuster's equation, the algorithmic model results with σ = 1.5 provide a better match than the results with a matching σ value. This is due to use of relative fitness score as opposed to absolute reproductive rate. It is also notable that the Peak 1 σ = 2.1 curve decreases with respect to the Peak 1 σ= 1.5 curve at larger population sizes; preliminary data obtained before running the original experiment shows that the Peak 1 σ = 1.5 curve also decreases in the same manner but at population sizes larger than those shown in the graph. The area of interest and focus of the study is the smaller population sizes of 100 individuals or less. The σ value can be seen to have no effect until the population size exceeds 400, therefore the validation of the results of the algorithmic method against Nowak and Schuster's predictions still holds.

haploid population of 100 individuals of length 30, this is at a per-base mutation rate of approximately 1.85%. Figure 5.4(b) shows the number of generations taken to lose each peak at this mutation rate, for each of the 2,000 runs. It is an apparent reversal of Figure 5.4(a) but with 100% of the runs having lost peak 0 within 200 generations. The population has almost entirely lost the ability to localize to either peak.

5.4 Discussion

At high mutation rates, individuals with greater mutational robustness can outcompete those with a higher fitness. It had previously been suggested that population size has no general effect on the position of the critical mutation rate, at which there is a phase transition from survival-of-the-fittest to survival-of-the-flattest [1]. However, the results of the current study suggest that population size does have an effect on the size of mutation rate that can be tolerated before the population loses the fittest and the flattest peaks, and that this is particularly noticeable in populations with 100 individuals or less; as shown in Figure 5.1, the size of mutation rate at which each peak is lost for increasing population sizes can be approximated by an exponential function. One possible reason for this is that small populations are more susceptible to stochastic variation due to random genetic drift [1, 54]; small populations with relatively large genomes cannot explore the entire neutral space of the landscape. Consequently, quasispecies formation is difficult, and the fitness peaks may be more easily lost. The dramatic reduction in critical mutation rate observed for small populations has implications for local extinction events in which there is a significant drop in population size; further work will be necessary to apply this result to populations under threat of local extinction. The null hypothesis (that critical mutation rate has no dependence on population size in haploid populations) can therefore be rejected. It can also be

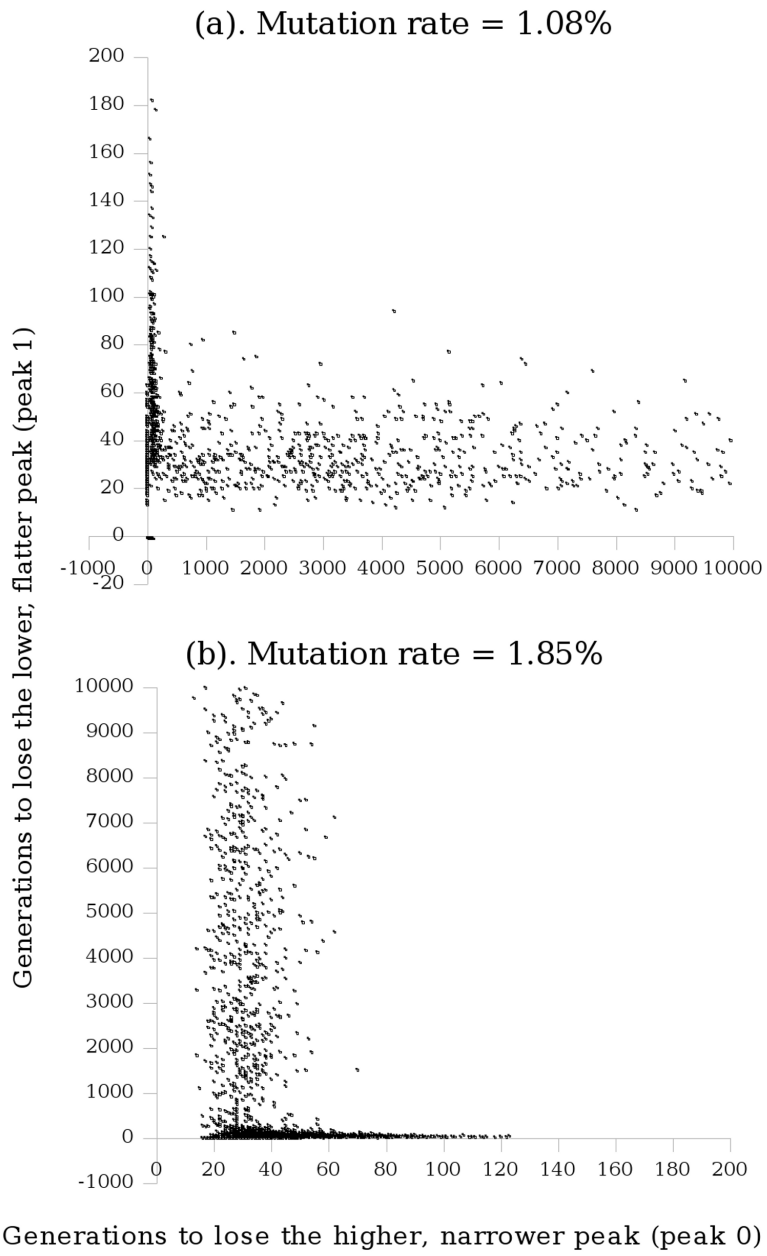

Figure 5.4: **Transition from survival-of-the-fittest to survival-of-the-flattest and subsequently to the error catastrophe.** Each point represents the number of generations it took to lose the high, narrow peak (peak 0) and the number to lose the lower, flatter peak (peak 1), in a single run of the GA for population size 100, sequence length 30. Where a peak was not lost within 10,000 generations, a value of -1 was assigned for that particular run of the genetic algorithm: all points on the negative side of either axis should be taken to have a value greater than 10,000. The critical mutation rate at which 95% of the runs had lost peak 0 (a) marks a point at which the population tends to converge on peak 1 due to survival-of-the-flattest. The error threshold at which 95% of the runs had lost peak 1 (b) marks a point at which mutation is too frequent for the population to maintain either peak.

observed that the curve obtained for the critical mutation rate flattens out to a greater degree than the curve obtained for the error threshold, as can be seen by looking at the faint lines in Figure 5.1. It is also noticeable by the difference in the value of the C parameter defined in Figure 5.1's caption, where $C = 1.4 \pm 0.2$ for the critical mutation rate and $C = 7.7 \pm 0.3$ for the error threshold; the lower the value of C, the flatter the curve. This explains why previous studies of larger populations have concluded that there is no relationship between the critical mutation rate and population size (e.g., [1]).

Verification of this method has been done using equations from analytical models (Equations 5.1 and 5.2) to produce comparable curves (Figure 5.2). Nowak and Schuster [32] present an analytical expression for the population size dependence of the error threshold using a system based on the Moran process (Equation 5.1). In Nowak and Schuster's system there is no crossover; population mixing is instead achieved by calculating transition probabilities based on the number of individuals that are a certain Hamming distance away from the master sequence (see section 4.2.2.2). This is comparable to our algorithmic method which introduces mixing through the biologically realistic process of crossover. Ochoa et al. [119, 80] include a reformulation of the Nowak and Schuster analytical expression (Equation 5.2), in which they make explicit the reduction in the error threshold when moving from infinite populations to those of size N. The observed consistency with the analytical error threshold models provides verification for our critical mutation rate results and our algorithmic method as a whole.

Previous studies have defined the critical mutation rate to be the midpoint between the highest mutation rate at which there is survival-of-the-fittest, and the lowest mutation rate at which there is survival-of-the-flattest [33, 1]. However, the results of

this study clearly show that there is a transition period from survival-of-the-fittest to survival-of-the-flattest (Figure 5.4), suggesting that a critical rate to consider is the mutation rate at which this transition begins; this was taken to be the highest mutation rate at which the high, narrow peak had been lost within the duration of the simulation in approximately 95% of cases. This defines the critical mutation rate to be at a point where survival-of-the-fittest is still the predominant outcome, but the population is no longer able to maintain the fittest peak indefinitely.

5.5 Chapter Summary

Research question:

- Does the critical mutation rate ever vary with population size?

Novel results:

- Critical mutation rate has an exponential dependence on population size in small populations.
- Justification for defining the critical mutation rate at a midpoint in the transition from survival-of-the-fittest to survival-of-the-flattest.

It was expected that population size should influence the size of mutation rate that can be tolerated before fitter individuals are outcompeted by those that have a greater mutational robustness, although previous studies had suggested this may not be the case [1]. The results of this study show that the size of mutation rate at which the high, narrow peak and the lower, flatter peak are lost for increasing population sizes can be approximated by an exponential function. The effect of population size on the size of mutation rate that can be tolerated before the population loses the fittest and the

flattest peaks is particularly noticeable in small populations with 100 individuals or less. This provides new insight into the factors that can affect survival-of-the-flattest in small populations, and has implications for populations under threat of local extinction. In addition, there is clear evidence for a range of mutation rates representing a transition period from survival-of-the-fittest to survival-of-the-flattest. This identifies a critical mutation rate representing the start of the transition period, which is defined as the highest mutation rate where survival-of-the-fittest is still the predominant outcome, but where the population is no longer able to maintain the fittest peak indefinitely.

Chapter 6

Critical Mutation Rates in Diploid Populations

6.1 Background and Hypothesis

All mammals have two copies of their genome; they are diploid as opposed to haploid [120]. In diploid organisms, one copy of the genome is inherited from the mother, while the other is inherited from the father. Each individual will therefore have two copies of each gene, each of which may be of a different form (a different allele). Different alleles have different degrees of dominance; an individual with two different alleles will display the phenotype of the dominant allele. In the majority of cases, mutant alleles are recessive while the non-mutant wild-type alleles are dominant [45].

Alves and Fontanari studied the error threshold in a single-peak fitness landscape with a diploid population [121]. The quasispecies model was used, in which a molecule is represented as a string of v digits, each of which is allowed to be one of k different values representing the different types of monomer used to make the molecule. The k^v different strings can be considered as different alleles of a gene that determines the fitness of a

haploid individual; this closely follows the classical one-locus, multiple-allele model of population genetics [54]. A diploid analogue of the single-peak fitness landscape was used, in accordance with the quasispecies model which was generalized by Wiehe et al. [38] to consider diploid individuals. There is a dominance parameter $-\infty < h < \infty$, where the master allele is completely dominant for $h = 1$ and completely recessive for $h = 0$ [121]. If $h = \frac{1}{2}$, there is no dominance. In addition to this, $h > 1$ models the case where there is heterozygote advantage, while $h < 0$ models heterozygote disadvantage. It was observed that, for $h \leqslant h_c \approx 1.75$, there are two distinct regimes: the quasispecies regime in which there is a single master allele around which most of the population is situated in sequence space, and the uniform regime where the 2^v alleles appear in the same proportion. They define the error threshold as being the error rate at which the transition between these two regimes occurs, with h_c representing a critical value beyond which the two regimes can no longer be distinguished. Beyond the error threshold the system undergoes an error catastrophe, something which was found to be postponed or even avoided in the case of a dominant allele ($h > \frac{1}{2}$). Based on the presence of an error threshold for a diploid population as described by Alves and Fontanari [121], it is expected that the relationship between population size and critical mutation rate observed for a haploid population should be conserved to some degree when moving from haploidy to diploidy:

Hypothesis - Critical mutation rate has a dependence on population size in diploid populations.

As diploid individuals have two copies of each sequence, this may confer a greater degree of robustness as any deleterious mutation will be potentially cancelled out; the second sequence has the potential to provide a back-up copy. This increased robustness may allow diploids to withstand higher mutation rates, and therefore have higher critical

mutation rates and error thresholds than haploids.

Ochoa and Jaffe [122] suggest there is an interaction between mutation rates and mating strategies in nature. Jacobi and Nordahl [123] performed computer simulations in which they introduced uniform crossover to a mutation model on an isolated-peak landscape. They found that in both the mutation and mutation with recombination models there is a phase transition from a localized to a non-localized state as the mutation rate is increased, and that this phase transition occurs at lower mutation rates with recombination. Boerlijst et al. [124] found that recombination lowers the mutation rate at which the error threshold occurs. Haploid systems use between-individual recombination while diploid systems use within-individual recombination:

Hypothesis - The magnitude of the critical mutation rate and error threshold will change when moving from haploidy to diploidy.

6.2 Methods

The hypotheses were tested using the diploid method described in section 4.2.2.3. As per chapter 5, the simulation was run for 10,000 generations, and the first generation at which there were no individuals on peak 0 was recorded. At this point, all the individuals were 2 or more mutations away from the top of peak 0 and therefore peak 0 was considered to have been lost (Figure 4.1). If peak loss did not occur within the 10,000 generations, a value of -1 was recorded in place of the generation number. Similarly, the number of generations it took to lose peak 1 was also recorded, where peak 1 was considered to have been lost when all individuals were 5 or more mutations away from the top of peak 1 (Figure 4.1). A range of per-base mutation rates was tested for a range of population sizes, with the simulation being repeated and run 2000 times for each combination. The mutation rate by which 95% of the runs had lost each

peak was recorded as a critical mutation rate, where a peak was considered to have not ever been lost only if there were individuals remaining on it at the end of the 10,000 generations. A range of values for the dominance parameter λ was tested, and the fitness of each individual was evaluated based on its constituent maternal and paternal sequences as per section 4.2.2.4. The experiment was repeated using differing types of recombination: haploid (between-individual recombination), diploid (within-individual recombination), diploid where the maternal sequence of one parent recombines with the maternal sequence of the other parent, and the paternal of one recombines with the paternal of the other (diploid but with haploid-like between-individual recombination), haploid with the recombination step omitted, and diploid with the recombination step omitted.

6.3 Results

6.3.1 The Relationship Between Critical Mutation Rate and Population Size is Conserved when Moving from Haploidy to Diploidy

Using a population of haploid individuals and a genetic algorithm with a simple two-peak fitness landscape (Figure 4.1), it was found that the mutation rates at which the high, narrow peak and the lower, flatter peak are lost (the survival-of-the-fittest and survival-of-the-flattest regimes ending at the critical mutation rate and error threshold respectively) for increasing population sizes could be approximated by an exponential function (see chapter 5 and Figure 5.1). This is conserved when moving to a population of diploid individuals, as can be seen in Figure 6.1. The dominance parameter is represented by λ (see section 4.2.2.4). It can be seen that as λ increases from 0.5 to 0.8,

the curve in the left-hand graph gets higher when population size exceeds approximately 200, while λ=0.8, 0.9 and 0.99 are all approximately in the same position. In the right-hand graph, there does not appear to be an affect when varying λ. The exception in both cases is λ=1.0. This is believed to be due to the method of calculating the overall fitness of each individual. As described in section 4.2.2.4, when λ<1.0, the fitness of both an individual's constituent sequences is taken into account to calculate an overall fitness score. However, when λ=1.0, only the fitness score of the fittest of the two constituent sequences will be taken into account. This means that the other constituent sequence could be anywhere in the fitness landscape; if the other sequence has a very low fitness, this will not be reflected in the overall fitness score meaning a low fitness sequence could be passed on to the next generation through tournament selection of a sequence with a high overall fitness score. Inclusion of λ=0.99 demonstrates this as a valid explanation as can be seen in Figure 6.1; the curve for λ=0.99 in both cases follows the apparent pattern of the other curves where λ=1.0 does not.

Transition between the states shown in Figure 5.4 is maintained when moving from haploidy to diploidy. Visualizing the relationship between population size, mutation rate and percentage of runs losing each peak shows the continuous transition from survival-of-the-fittest to survival-of-the-flattest (around the critical mutation rate) and subsequently to the error catastrophe (around the error threshold), and emphasizes the relationship between these transitions (Figure 6.2). For example, for population sizes of several hundred individuals, the lower dashed line across the lower projections in Figure 6.2 indicates approximately where the percentage loss of peak 0 begins to rise steeply and that of peak 1 begins to fall steeply as mutation rate is increased: the transition from survival-of-the-fittest to survival-of-the-flattest; and the upper dashed line indicates approximately where the percentage loss of peak 0 has reached 100% and

that of peak 1 has reached its minimum before rising back upward as mutation rate is increased further: the transition from survival-of-the-flattest to the error catastrophe. In the upper projection (b) of Figure 6.2 it can be seen that for smaller population sizes (less than 50) the percentage of runs losing peak 1 does not fall below approximately 70%. This suggests 70% loss of peak 1 as a lower bound when considering error threshold. Below 50% loss of peak 0, individuals have transferred from peak 1 to peak 0, so 50% is a lower bound for considering critical mutation rate. The shapes of the population size to mutation rate mappings become increasingly consistent as these lower bounds are exceeded and 95% peak loss is a good choice for both critical mutation rate and error threshold.

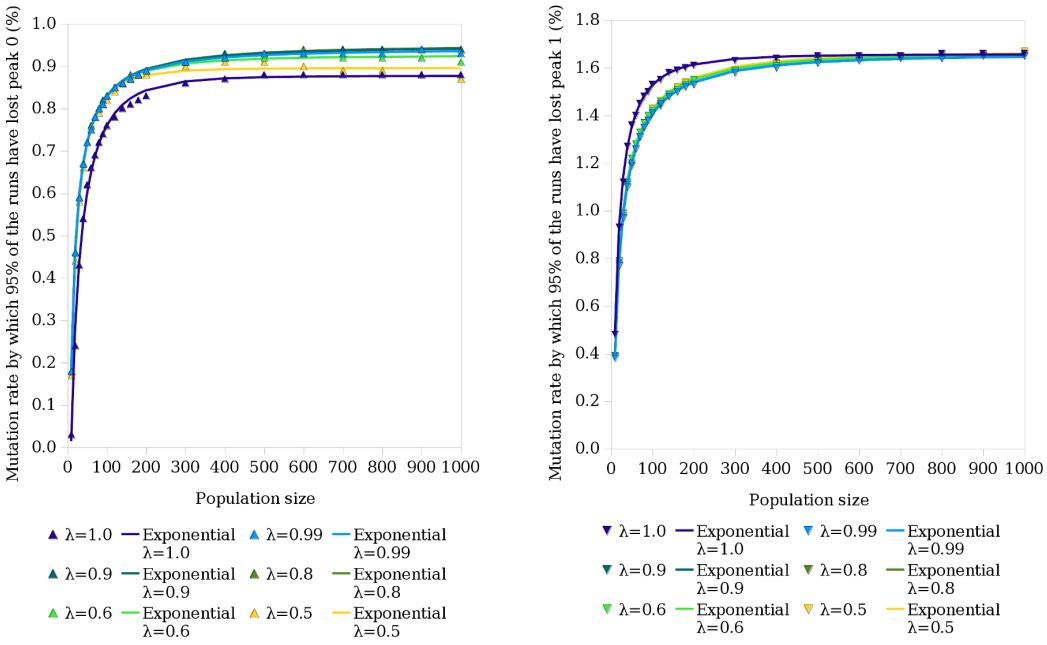

Figure 6.1: **Critical mutation rate has an exponential dependence on population size in diploids.** Here λ is the dominance parameter, as described in the section entitled Fitness Calculation. The simulation was run using the λ values listed. The points show the results obtained, which can be approximated by exponential functions as shown by the lines (obtained by curve-fitting using R with a least squares method). It should be noted that, in the left graph, λ=0.8 appears directly underneath λ=0.9, while in the right graph both λ=0.8 and λ=0.9 overlap with λ=0.5, 0.6, and 0.99. The left graph shows the curve obtained for the critical mutation rate and the right graph shows the error threshold, both for a diploid population. Refer to Figure 5.1 for the equivalent curves for a haploid population.

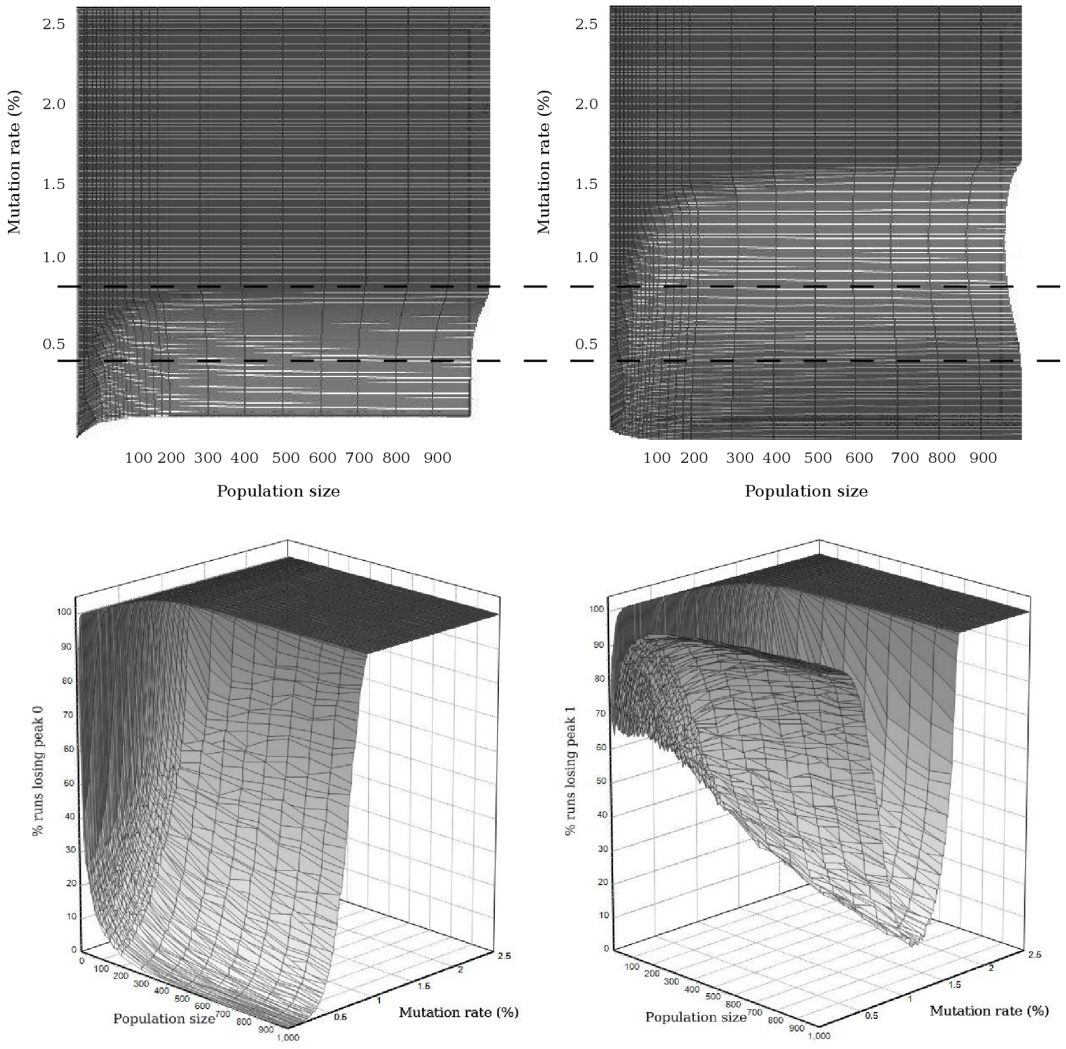

Figure 6.2: **Percentage of runs losing the peaks at different mutation rates and population sizes.** The results shown are for the diploid method with $\lambda = 1.0$, for peak 0 (a, left) and peak 1 (b, right). In the two upper projections the axis coming out of the page is the percentage of runs. The lower dashed line across these projections indicates the transition from survival-of-the-fittest to survival-of-the-flattest. The upper dashed line indicates the transition from survival-of-the-flattest to the error catastrophe.

6.3.2 Haploid and Diploid Recombination Systems Affect the Critical Mutation Rate and Error Threshold

Section 6.3.1 confirmed that the result using the haploid system also applies to a diploid population. However, the diploid critical mutation rate and error threshold curves are lower than those for a haploid population. This means the null hypothesis (that the magnitude of the critical mutation rate and error threshold will not change when moving from haploidy to diploidy) can be rejected. The experiment was therefore repeated using differing types of recombination: haploid (between-individual recombination), diploid (within-individual recombination), diploid where the maternal sequence of one parent recombines with the maternal sequence of the other parent, and the paternal of one recombines with the paternal of the other (diploid but with haploid-like between-individual recombination), haploid with the recombination step omitted, and diploid with the recombination step omitted. The results confirm that varying the type of recombination varies the critical mutation rate curve (Figure 6.3).

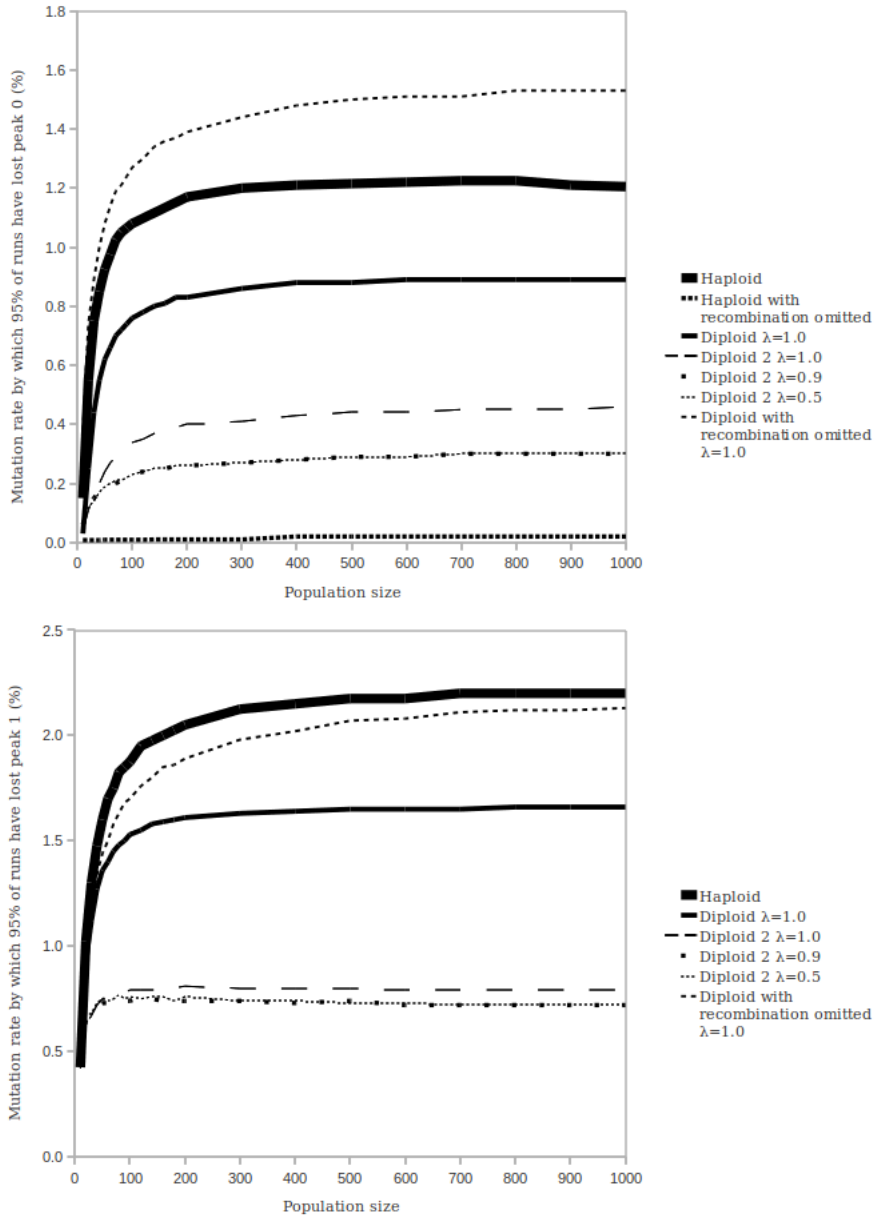

Figure 6.3: **Comparison of mutation rate curves for systems with different types of recombination.** For haploid (single-sequenced individuals), parent A recombines with parent B. For diploid, the individual's maternal sequence recombines with it's paternal sequence. Diploid 2 has a modified version of recombination where the maternal sequence of one parent recombines with the maternal sequence of the other parent (and likewise with the paternal sequences). In the case of diploid with recombination omitted, the child consists of the maternal sequence of one parent, and the paternal sequence of the other parent. It should be noted that haploid with recombination omitted is not displayed on the lower graph as Peak 1 was never lost.

6.4 Discussion

Based on the observation that the error threshold has a dependence on haploid population size, and the observation by Wiehe et al. [38] that this relationship is not lost in diploid systems, a hypothesis was formed that the relationship will also hold for the critical mutation rate in haploid and diploid systems with a two-peak landscape. In a diploid system modelled on the process of meiosis in biology, each individual has two copies of the genetic sequence and recombination occurs within as opposed to between individuals. The resulting single-sequenced gamete then joins with another to form a child. The haploid and diploid methods of reproduction are fundamentally different; single-sequence versus two-sequence individuals, and between-individual recombination versus within-individual recombination means two populations reproducing using the two different systems will differ in their occupation of sequence space. The two copies of each sequence present in diploid individuals also gives them a redundancy not found in haploids. It was therefore expected that there would be some variation in the results when the experiments with a haploid system were reproduced using a diploid system. Consistent with this, the results using the haploid system also apply to a diploid population, but the diploid critical mutation rate and error threshold curves are lower than those for a haploid population (Figures 5.1, 6.1 and 6.2). The null hypothesis (that critical mutation rate has no dependence on population size in diploid populations) can therefore be rejected.

It has been suggested that there is an interaction between mutation rates and mating strategies in nature [122]. Haploid systems use between-individual recombination while diploid systems use within-individual recombination. Recombination lowers the mutation rate at which the error threshold occurs [124]. Assortative, non-random mating, in which individuals of a similar phenotype mate more often than expected by

chance, is able to overcome this shift toward lower error threshold magnitudes induced by recombination [122]. Conversely, dissortative mating, in which dissimilar individuals mate more often, reduces the magnitude of the error threshold. In the haploid system, the simulation starts with the population clustered at the two peaks. As the simulation is run, the population tends towards one of the peaks assuming the mutation rate does not exceed the error threshold. Recombination therefore tends to occur between sequences with similar fitnesses, and mating can be considered to be assortative. In our diploid system, the simulation starts with the population clustered at the two peaks, with individuals either completely at either peak, or with one sequence at one peak and one at the other. As fitness is calculated as a single value based on the fitness of an individual's two constituent sequences (see section 4.2.2.4), an individual can have, for example, a high fitness value but consist of two sequences in completely different parts of the fitness landscape. There is therefore a chance that the individuals selected to mate could have very different genetic make-ups; the degree of dissortative mating exceeds that of the haploid system. The difference in mating systems used by haploids and diploids is a potential reason for the difference in the curves shown in Figure 6.1; further work will be required to confirm this.

The effect of increasing λ on the critical mutation rate can be explained by the fact that lower λ values means lower overall fitness scores for each individual; lower λ values give more weight to the sequence with the lower fitness value than when λ is high (see section 4.2.2.4). This explains why lower λ values gave rise to critical mutation rate curves that flattened out to lower mutation rates. When λ was 0.8 or above, enough weight was given to the higher fitness sequence so that this effect no longer occured. In contrast the error threshold does not appear to be affected by changing the λ value. Approaching the error threshold (defined as the mutation rate at which 95%

of individuals have lost the lower peak), the majority of individuals will be occupying areas of the fitness landscape with lower fitness (after having subsequently exceeded the critical mutation and therefore no longer occupying the higher peak). This reduces the potential difference in fitness between the two constituent sequences of an individual, meaning calculation of the overall fitness score will be less affected by λ.

In chapter 5, an existing definition of the critical mutation rate was improved upon. Previous studies had defined it as the midpoint between the highest mutation rate at which there is survival-of-the-fittest, and the lowest mutation rate at which there is survival-of-the-flattest [33, 1]. The results of this study clearly show that there is a transition from survival-of-the-fittest to survival-of-the-flattest and subsequently to the error catastrophe (Figure 5.4). Figure 6.2 shows these transitions occurring in a diploid population, and demonstrates a relationship between the critical mutation rate and the error threshold. The highest point at lower mutation rates in (b) appears to correspond to where the curve in (a) starts to ascend. Likewise, by the time the curve in (b) has descended to its lowest, the curve in (a) has reached its highest. This shows the transition of the population favouring peak 0 to favouring peak 1. The transition occurs around the critical mutation rate. At less than 50% loss of peak 0, individuals are still moving from peak 1 to peak 0. The critical mutation rate concerns the loss of individuals from peak 0 to peak 1, therefore the critical mutation rate should not be considered to be at a point where there is still a significant transition in the other direction (implying there is still a peak 0 advantage). In the top graph in Figure 6.2 (b), it can be seen that for smaller population sizes (less than 50), the curve does not fall below approximately 70% loss of peak 1. Considering the equivalent portion of the graph, Figure 6.2 (a) suggests that considering a peak loss of anything much less than 50% will be redundant when the population is small. The critical mutation rate should

be considered not as a single value at the midpoint, but rather as *lying within a range of values with a lower limit of 50% loss of the high, narrow peak.*

The critical mutation rate curves observed for a diploid system are lower than those observed for a haploid system (Figure 6.1). Ochoa and Jaffe [122] suggested there is an interaction between mutation rates and mating strategies in nature. Haploid systems use between-individual recombination while diploid systems use within-individual recombination; the results confirm that varying the type of recombination varies the critical mutation rate curve, and suggests this as a reason for the difference in curves observed for haploid and diploid systems (Figure 6.3). Recombination lowers the mutation rate at which the error threshold occurs [124, 9]. Ochoa and Jaffe [122] suggest that assortative, non-random mating, in which individuals of a similar phenotype mate more often than expected by chance, is able to overcome this shift toward lower error threshold magnitudes induced by recombination. Converseley, dissortative mating, in which dissimilar individuals mate more often, reduces the magnitude of the error threshold. In the haploid system described here and in [41], the simulation starts with the population clustered at the two peaks. As the simulation is run, the population tends towards one of the peaks assuming the mutation rate does not exceed the error threshold. Recombination therefore tends to occur between sequences with similar fitnesses, and mating can be considered to be assortative. In the diploid system described for the current study, the simulation starts with the population clustered at the two peaks, with individuals either completely at either peak, or with one sequence at one peak and one at the other. As fitness is calculated as a single value based on the fitness of an individual's two constituent sequences (and a given λ value), an individual can have, for example, a high fitness value but consist of two sequences in completely different parts of the fitness landscape. There is therefore a chance that the individuals

selected to mate could have very different genetic make-ups; the degree of dissortative mating exceeds that of the haploid system.

A modified diploid system was designed as an intermediate between the recombination strategies of haploid and diploid systems (between- and within-individual recombination respectively). This hybrid system was denoted 'diploid 2', in which the maternal sequence of one parent recombines with the maternal sequence of the other parent (and likewise with the paternal sequences). It was expected that the mutation curve obtained for diploid 2 would be somewhere between the curves obtained for haploid and diploid. However, the mutation curve for diploid 2 was found to be lower than both the haploid and diploid curves (Figure 6.3). Consideration of recombination in the diploid and diploid 2 systems provides a possible explanation for this result. Assume a simple case in which there are two parent individuals (1 and 2), each with a single deleterious mutation (δ_1 and δ_2) on their maternal strand, and that λ=1.0. In the diploid system, the maternal and paternal strands of an individual recombine to produce a single-stranded gamete. There is a 1 in 2 chance the gamete for parent 1 will have δ_1 (and likewise for parent 2). There is therefore a 1 in 4 chance that when the two gametes come together to form the child individual, the child will inherit neither mutation, and a 3 in 4 chance it will inherit either δ_1 or δ_2 or both. This translates to a 1 in 4 chance that the child will inherit both δ_1 and δ_2, and a 1 in 4 chance it will inherit just one of the mutations. In the diploid 2 system, assuming the same parent individuals, the recombination will occur between the maternal strand of parent 1 and the maternal strand of parent 2, each carrying δ_1 and δ_2 respectively. There is a 1 in 2 chance the resulting gamete will inherit δ_1, and a 1 in 2 chance it will inherit δ_2. However, there are no mutations on the paternal strands, and so the gamete produced as a result of their recombination will inherit no mutations. This means that the child

in the diploid 2 system will have one strand with potentially two mutations, and one strand completely free of mutations; even if both mutations are inherited on the maternal strand, they will be masked by the paternal strand during the calculation of fitness. In subsequent generations, the fitness of the individual could potentially drop from f to $f - x$, where x may be a large number if many previously masked deleterious mutations have been inherited. The diploid system does not mask the mutations, therefore fitness is more likely to drop gradually rather than abruptly. This suggests a greater robustness in the diploid system. This simple case demonstrates the potential difference between the within- and between-individual recombination systems, and suggests greater robustness as the reason the diploid system can withstand higher mutation rates than the diploid 2 system.

6.5 Chapter Summary

Research questions:

- Does the critical mutation rate ever vary with population size in a diploid population replicating using a system modelled on the biological process of meiosis?
- Do the haploid and diploid critical mutation rate curves vary due to the differing types of recombination in each respective system?

Novel results:

- Extension of the model to include diploidy, demonstrating that the critical mutation rate still has an exponential dependence on population size when using a genetic algorithm that is diploid and modelled closely on a real biological system.
- Other studies have suggested different types of recombination affect the error threshold. Extending from this, it has been shown that critical mutation rates are

lower for diploid populations than haploid populations because of the difference in recombination.

- Further justification for defining the critical mutation rate at a midpoint in the transition from survival-of-the-fittest to survival-of-the-flattest.

Chapter 7

Critical Mutation Rates in Real Biological Systems

It has been established in Chapters 5 and 6 that critical mutation rate has an exponential dependence on population size in both haploid and diploid populations. This result was obtained for a population of individuals of sequence length 30 evolving on a two-peak landscape. While this result provides novel insight into the factors that can influence the fitness of a population, an important question to ask is: is this result relevant to real biological systems? More specifically, is the exponential model relevant to natural as well as artificial populations? In naturally occurring species there are many factors that can influence gene expression such as interactions between genes through epistasis, the presence of non-coding segments of DNA indirectly involved in gene expression through binding regulatory proteins, and neutrality, all of which can vary depending on the species.

The system described in section 4.2.2.3 developed the haploid genetic algorithm by not only introducing diploidy, but also through modelling the algorithm on the biological process of meiosis in an attempt to make the first step towards bridging

the gap between artificial and biological evolution. While biological systems can be complex, one of the next steps is to bring an artificial system into the realm of biology by using parameter values within the range observed in nature.

7.1 From Artificial to Biological Evolution: Mutation of Genes in Nature

A review of the literature and the BioNumbers database [125] provided a range of observed values for critical evolutionary parameters. The focus of the study is the critical mutation rate, therefore mutation rate was selected as one of the parameters to be tested within a biological range. The range is given for various species in Table 7.1, all taken from the BioNumbers database (with the exception of the genome size for the sheep/cow line which was taken from [126]). Other parameters were gene length and the Hamming distance between gene variants (alleles). Assuming peak 0 and peak 1 represent two different alleles, the latter parameter gives the distance between the two peaks. Recombination was limited to one event per replication and selection proceeded as described in the simulation model, as the rate or strength of either of these respectively was not the focus of the study. It should be noted that assuming equal strength of selection should not be expected to affect the credibility of the results; Whitlock et al. [127] performed computer simulations to investigate the effects of varying the strength of selection and mutational effects among dimensions. They used a model based on Fisher's model of the geometry of adaptation [45], but used a hyperellipse in which the strength of selection along any axis was drawn from an exponential distribution. They concluded that changing from a hypersphere to a hyperellipse, and thus introducing dimensions with stronger selection than others, had

a negligible effect on their results.

7.1.1 Mutation Rates

Table 7.1 lists mutation rates obtained from various listed sources, categorized by whether they refer to pre or post selection rates, and whether they are the rates per base or per genome. As the models described in section 4 are effectively non neutral in that the fitness of the individual is inferred directly from the sequence, and selection is done based on the relative fitnesses of the individuals, the mutation rate used in the simulation model is analagous to the biological pre selection, per base mutation rate. Genome lengths are included for each species for reference when considering the per genome mutation rate. Nachman and Crowell [128] obtained an estimate of the average mutation rate per nucleotide by comparing pseudogenes (genes that do not code for proteins or are never expressed) in humans and chimpanzees. Baer et al. [129] brought together the results of a number of studies, both theoretical and empirical, to list mutation rate estimates in a number of multicellular eukaryotes. Drake et al. [130] list mutation rate estimates for species based on a number of studies, including mutation accumulation and radiation experiments. Lynch [131] also lists mutation rates from various sources. Xue et al. [132] obtained an estimate for the base substitution rate in the human Y chromosome through direct sequencing. Kumar and Subramanian [133] conducted a computational analysis of 5669 genes from species of placental mammals. Keightley et al. [134] did whole-genome shotgun sequencing of three mutation accumulation lines of *Drosophila melanogaster*. Denver et al. [135] provide a direct estimate of the mutation rate from a set of *Caenorhabditis elegans* mutation accumulation lines. Haag-Liautard et al. [136] and Ossowski et al. [137] provide estimates using mutation accumulation lines. Durbin et al. [138] examined

Table 7.1: **Mutation rates for various species.**

Source	*Species*	*Genome size (bp)*	*Mutation rate*	*Selection*	*Base/genome*	*Other*
Nachman and Crowell (2000)[128]	Human	3080000000	2.50E-008	Pre	Per base	Per generation
Nachman and Crowell (2000)[128]	Human	3080000000	1.75E+002	Pre	Per genome	Per generation
Baer et al. (2007)[129]	*Drosophila melanogaster*	120000000	1.00E+000	Pre	Per genome	Diploid
Baer et al. (2007)[129]	*Drosophila* spp.	120000000	7.00E-002	Pre	Per genome	Diploid
Baer et al. (2007)[129]	*Drosophila melanogaster*	120000000	1.00E-001	Pre	Per genome	Diploid
Baer et al. (2007)[129]	*Drosophila melanogaster*	120000000	1.20E+000	Pre	Per genome	Diploid
Baer et al. (2007)[129]	Quail, chicken	1050000000	4.90E-001	Pre	Per genome	Diploid
Baer et al. (2007)[129]	Sheep, cow	2870000000	9.00E-001	Pre	Per genome	Diploid
Baer et al. (2007)[129]	Old World Monkey		1.90E+000	Pre	Per genome	Diploid
Baer et al. (2007)[129]	Mouse, rat	2640000000	9.10E-001	Pre	Per genome	Diploid
Baer et al. (2007)[129]	Human, chimpanzee	3080000000	3.00E+000	Pre	Per genome	Diploid
Drake et al. (1998)[130]	*Drosophila melanogaster*	120000000	3.40E-010	Post	Per base	Per replication
Drake et al. (1998)[130]	Mouse	2640000000	1.80E-010	Post	Per base	Per cell division
Drake et al. (1998)[130]	Mouse	2640000000	1.10E-008	Post	Per base	Per generation
Drake et al. (1998)[130]	Human	3080000000	5.00E-011	Post	Per base	Per replication
Drake et al. (1998)[130]	Human	3080000000	1.60E-001	Post	Per genome	Per replication
Lynch (2010)[131]	Human	3080000000	1.29E-008	Post	Per base	Per generation
Lynch (2010)[131]	Human	3080000000	6.00E-002	Post	Per base	Per cell division
Lynch (2010)[131]	*Drosophila melanogaster*	120000000	4.65E-009	Pre	Per base	Per generation
Lynch (2010)[131]	*Drosophila melanogaster*	120000000	1.30E-010	Pre	Per base	Per cell division
Lynch (2010)[131]	*Saccharomyces cerevisiae*	12100000	3.30E-010	Pre	Per base	Per generation
Xue et al. (2009)[132]	Human (Y chromosome)	58000000	3.00E-008	Pre	Per base	Per generation
Kumar and Subramanian (2002)[133]	Average mammalian		2.20E-009	Pre	Per base	Per genome/year
Kumar and Subramanian (2002)[133]	Mammalian upper bound		2.61E-009	Pre	Per base	Per genome/year
Keightley et al. (2009)[134]	*Drosophila melanogaster*	120000000	3.46E-009	Pre	Per base	Per generation
Denver et al. (2004)[135]	*Caenorhabditis elegans*	100000000	2.10E-008	Pre	Per base	Per generation
Haag-Liautard et al. (2008)[136]	*Drosophila melanogaster*	120000000	9.90E-001	Pre	Per genome	Per generation
Haag-Liautard et al. (2008)[136]	*Caenorhabditis elegans*	100000000	8.40E-009	Pre	Per base	Per generation
Haag-Liautard et al. (2008)[136]	*Drosophila melanogaster*	120000000	6.20E-008	Pre	Per base	Per generation
Ossowski et al. (2010)[137]	*Arabidopsis thaliana*	157000000	7.10E-009	Pre	Per base	Per generation
Durbin et al. (2010)[138]	Human	3080000000	1.00E-008	Pre	Per base	Per generation
Lynch et al. (2008)[8]	*Caenorhabditis elegans*	100000000	2.90E+000	Pre	Per genome	Per generation
Lynch et al. (2008)[8]	*Saccharomyces cerevisiae*	12100000	3.30E-010	Pre	Per base	Per cell division
Lynch (2010) [139]	*Drosophila melanogaster*	120000000	4.65E-009	Post	Per base	Per generation
Lynch (2010) [139]	*Caenorhabditis elegans*	100000000	5.60E-009	Post	Per base	Per generation
Lynch (2010) [139]	*Arabidopsis thaliana*	157000000	6.50E-009	Post	Per base	Per generation
Lynch (2010) [139]	*Saccharomyces cerevisiae*	12100000	3.30E-010	Post	Per base	Per generation
Lynch (2010) [139]	Human	3080000000	1.28E-008	Post	Per base	Per generation

variation in the sequence of the human genome. Lynch et al. [8] provide a mutation rate estimate from complete genome sequencing of *Saccharomyces cerevisiae.* Lynch [139] used existing data to estimate the mutation rate of various eukaryotes.

7.1.2 Genetic Sequences

Tables 7.2 and 7.3 list the length of various genes and the approximate length of sequences calculated from the genome size and gene number respectively. It should be noted that Table 7.3 was intended to give an estimate of appropriate sequence lengths to use in the simulation. It does not take into account the proportion of each genome comprising introns and exons and therefore should not be considered to be a direct estimate of average gene length. If each peak in the two-peak landscape is considered to be a different allele (variant of a gene), the values listed in Table 7.4 can be seen to be analogous to the distance between the peaks. Similarly, Table 7.5 lists the number of differences (polymorphisms) between human genes which may also be analogous to the distance between the peaks if we consider the peaks to represent two different genes. In both cases the value of 10 used in the haploid and diploid models to produce the results presented in Chapters 5 and 6 is close to the range of numbers listed therefore it was decided to keep this number constant. Varying the distance between the peaks may be an interesting future study.

As can be seen from Tables 7.2 and 7.3, the sequence length of 30 used to produce the results in Chapters 5 and 6 is small when compared with the length of genes found in a wide range of natural species. While the diploid method described in section 4.2.2.3 was designed so that different parameter values could be input with each run, the process of mutating each sequence base by base means that increasing the sequence length has a significant impact on the overall runtime. The large range of sequence

Table 7.2: **Gene lengths for various species.**

Source	*Species*	*Gene*	*Length*	*Units*
Derelle et al. (2006)[140]	*Arabidopsis thaliana*	Mean gene size	2232	bp
Derelle et al. (2006)[140]	*Schizosaccharomyces pombe*	Mean gene size	1426	bp
Sharma et al. (2005)[141]	Human	Histone proteins family	3.55	kbp
Sharma et al. (2005)[141]	Human	Interleukins	14.67	kbp
Sharma et al. (2005)[141]	Human	Serine (or cysteine) proteinase inhibitor family	17.87	kbp
Sharma et al. (2005)[141]	Human	Tumor necrosis factor (ligand) superfamily	23.49	kbp
Sharma et al. (2005)[141]	Human	CD antigens	26.96	kbp
Sharma et al. (2005)[141]	Human	Immunoglobulin heavy chains	0.38	kbp
Sharma et al. (2005)[141]	Human	Immunoglobulin kappa chains	0.55	kbp
Sharma et al. (2005)[141]	Human	Immunoglobulin lambda chains	0.35	kbp
Sharma et al. (2005)[141]	Human	Interleukin receptors family	29.77	kbp
Sharma et al. (2005)[141]	Human	T cell receptor beta chains	0.42	kbp
Sharma et al. (2005)[141]	Human	Homeo box	5.48	kbp
Sharma et al. (2005)[141]	Human	Eukaryotic transcription factor	36.54	kbp
Sharma et al. (2005)[141]	Human	Zinc finger protein family	30.33	kbp
Sharma et al. (2005)[141]	Human	DEAD/H (Asp-Glu-Ala-Asp/His) box polypeptides	43.44	kbp
Sharma et al. (2005)[141]	Human	Ribosomal protein genes	4.94	kbp
Sharma et al. (2005)[141]	Human	Mitochondrial ribosomal protein genes	16.92	kbp
Sharma et al. (2005)[141]	Human	Cytochrome P450 superfamily	31.34	kbp
Sharma et al. (2005)[141]	Human	Proteasome subunit genes	25.64	kbp
Sharma et al. (2005)[141]	Human	G protein-coupled receptor family	24.71	kbp
Sharma et al. (2005)[141]	Human	Tripartite motif-containing family	29.29	kbp
Sharma et al. (2005)[141]	Human	Solute carrier family	59.19	kbp
Sharma et al. (2005)[141]	Human	RAS oncogene family	39.92	kbp
Sharma et al. (2005)[141]	Human	ATP-binding cassette transporters gene family	73.85	kbp
Sharma et al. (2005)[141]	Human	G protein polypeptide genes	58.83	kbp
Sharma et al. (2005)[141]	Human	Potassium voltage-gated channel genes	104.95	kbp
Sharma et al. (2005)[141]	Human	Protein phosphatase subunit genes	65.62	kbp
Sharma et al. (2005)[141]	Human	Collagen family	132.83	kbp
B. Lewin, Genes IX (2008) pg 43+[142]	Yeast	Average gene length	1.4	kb
B. Lewin, Genes IX (2008) pg 43+[142]	Yeast	Maximum gene length (very few exceed this)	5	kb
B. Lewin, Genes IX (2008) pg 43+[142]	Human	Average gene length	27	kb
B. Lewin, Genes IX (2008) pg 43+[142]	Flies or mammals	Minimum gene length (very few shorter than this)	2	kb
B. Lewin, Genes IX (2008) pg 43+[142]	Flies or mammals	Usual gene length range	5 to 100	kb

Table 7.3: **Estimated sequence lengths based on genome size and gene number estimations.**

Source	*Species*	*Gene*	*Length*	*Units*
Xu et al. (2006) Supplementary Table 1 Eukaryotes[143]	*Anopheles gambiae* (mosquito)	Estimated genome size/gene number	20.34	kb
Xu et al. (2006) Supplementary Table 1 Eukaryotes[143]	*Arabidopsis thaliana* (plant)	Estimated genome size/gene number	4.53	kb
Xu et al. (2006) Supplementary Table 1 Eukaryotes[143]	*Ashbya gossypii* (ascomycete)	Estimated genome size/gene number	1.84	kb
Xu et al. (2006) Supplementary Table 1 Eukaryotes[143]	*Caenorhabditis elegans* (nematode)	Estimated genome size/gene number	5.02	kb
Xu et al. (2006) Supplementary Table 1 Eukaryotes[143]	*Ciona intestinalis* (sea squirt)	Estimated genome size/gene number	10	kb
Xu et al. (2006) Supplementary Table 1 Eukaryotes[143]	*Cyanidioschyzon merolae* (red alga)	Estimated genome size/gene number	3.1	kb
Xu et al. (2006) Supplementary Table 1 Eukaryotes[143]	*Dictyostelium discoideum* (amoeba)	Estimated genome size/gene number	2.82	kb
Xu et al. (2006) Supplementary Table 1 Eukaryotes[143]	*Drosophila melanogaster* (fruit fly)	Estimated genome size/gene number	8.82	kb
Xu et al. (2006) Supplementary Table 1 Eukaryotes[143]	*Fugu rubripes* (puffer fish)	Estimated genome size/gene number	11.41	kb
Xu et al. (2006) Supplementary Table 1 Eukaryotes[143]	Human	Estimated genome size/gene number	85.71	kb
Xu et al. (2006) Supplementary Table 1 Eukaryotes[143]	Mouse	Estimated genome size/gene number	71.43	kb
Xu et al. (2006) Supplementary Table 1 Eukaryotes[143]	*Neurospora crassa* (fungus)	Estimated genome size/gene number	3.97	kb
Xu et al. (2006) Supplementary Table 1 Eukaryotes[143]	*Oryza sativa* Chr. 1 (rice)	Estimated genome size/gene number	6.77	kb
Xu et al. (2006) Supplementary Table 1 Eukaryotes[143]	Rat	Estimated genome size/gene number	83.33	kb
Xu et al. (2006) Supplementary Table 1 Eukaryotes[143]	*Saccharomyces cerevisiae* (yeast)	Estimated genome size/gene number	2.11	kb
Xu et al. (2006) Supplementary Table 1 Eukaryotes[143]	Chicken	Estimated genome size/gene number	46.22	kb
Xu et al. (2006) Supplementary Table 1 Eukaryotes[143]	*Caenorhabditis briggsae* (nematode)	Estimated genome size/gene number	5.33	kb
Xu et al. (2006) Supplementary Table 1 Eukaryotes[143]	*Phanerochaete chrysosporium* (fungus)	Estimated genome size/gene number	2.55	kb
Xu et al. (2006) Supplementary Table 1 Eukaryotes[143]	*Tetraodon nigroviridis* (puffer fish)	Estimated genome size/gene number	12.18	kb

Table 7.4: **Genetic distance between alleles for various genes.**

Source	*Genetic distance between alleles*	
Sachs et al. (1986)[144]	2.02% difference	From Table 1: Sequence differences between Adh1-1F and Adh1-1S alleles
Bryan et al. (2000)[145]	1 amino acid	Rice blast resistance (R) gene Pi-ta, which encodes a 928-amino acid receptor. There are two alleles, the one that encodes the Pi-ta resistance protein, and the one that leaves rice susceptible. The susceptible version has serine instead of alanine at position 918
Ramkumar et al. (2010)[146]	13	SNPs in Pikh allele for different rice varieties (taken from Table 1)
	7	
	7	
	19	
	45	
	48	
	54	
	19	
	56	
	47	
Cummings and Clegg (1998)[147]	19	Base pairs different between Adh1 alleles of wild barley

Table 7.5: **Number of polymorphisms in various human genes taken from Table 1 in [3]**

Gene	No. synonymous polymorphisms	No. non-synonymous polymorphisms	No. non-coding polymorphisms
HMGCR	0	1	1
HSD3B1	3	2	1
HSD3B2	1	1	2
HTR1A	1	0	1
HTR1D	1	1	0
HTR1DB	2	0	1
HTR1E	1	1	0
HTR1EL	1	0	0
HTR2A	2	3	9
HTR2C	0	1	0
HTR5A	2	0	0
HTR6	1	0	0
HTR7	0	0	0
IGF1	0	0	8
IGF2	0	0	1
ITGA2B	4	3	0
ITGB3	4	3	0
KLK2	0	1	2
LCAT	3	0	0
LDLR	7	3	0
LIPC	4	3	4
LPL	1	1	0
MAOA	1	0	0
MAOB	1	0	0
MPL	1	2	1
NGFB	1	1	5
NOS1	0	0	0
NT3	1	0	0
NTRK1	5	2	0
PACE	2	0	4
PAI1	1	2	1
PAI2	5	4	5
PC1	1	3	1
PCI	5	5	4
POMC	0	0	0
PRL	1	1	1
PROC	3	0	0
PROS1	1	0	0
PTAFR	0	2	0
PTH	1	0	2
PTHLH	0	0	13
SELP	5	8	0
SHBG	1	3	1
SLC6A1	2	0	2
SLC6A3	6	1	0
SLC6A4	1	2	1
TBXA2R	1	0	0
TBXAS1	1	6	1
TFPI	0	1	0
TH	1	1	0
THBD	0	0	0
THPO	0	0	2
VLDLR	3	1	2
All genes	207	185	168

lengths in Tables 7.2 and 7.3 identified a need to improve the algorithm's efficiency by removing the scaling of sequence length and runtime. This was achieved by modifying the way the population is stored and mutated, and is described in section 4.2.2.5.

The modified diploid simulation method with improved efficiency was run using mutation rates between 1×10^{-10} and 0.01, sequence lengths between 2000 and 150000, the distance between the two peaks constant at 10, and with population sizes ranging from 10 to 1000 as before. Dominance was set to a fraction below 1.0. This was to account for the drop in critical mutation rate observed in Figure 6.1 when λ=1.0, This was believed to be due to the method of calculating the overall fitness of each individual. As described in section 4.2.2.4, when λ<1.0, the fitness of both an individual's constituent sequences is taken into account to calculate an overall fitness score. However, when λ=1.0, only the fitness score of the fittest of the two constituent sequences will be taken into account. This means that the other constituent sequence could be anywhere in the fitness landscape; if the other sequence has a very low fitness, this will not be reflected in the overall fitness score meaning a low fitness sequence could be passed on to the next generation through tournament selection of a sequence with a high overall fitness score. Mutation is generally expected to occur at rates below that of the error threshold, as discussed in section 3.1.5. It is expected that mutation will also typically occur at rates below the critical mutation rate as this will enable populations to climb the fittest peak. Longer sequences means more bases to potentially mutate each generation:

Hypothesis - Increasing the sequence length will lower both the critical mutation rate and error threshold in line with the exponential model. Neither the critical mutation rate nor the error threshold will go below the typical mutation rates found in nature.

7.2 Methods

The hypothesis was tested using the diploid method with improved efficiency described in section 4.2.2.5. As per Chapters 5 and 6, the simulation was run for 10,000 generations, and the first generation at which there were no individuals on peak 0 was recorded. At this point, all the individuals were 2 or more mutations away from the top of peak 0 and therefore peak 0 was considered to have been lost (Figure 4.1). If peak loss did not occur within the 10,000 generations, a value of -1 was recorded in place of the generation number. Similarly, the number of generations it took to lose peak 1 was also recorded, where peak 1 was considered to have been lost when all individuals were 5 or more mutations away from the top of peak 1 (Figure 4.1). A range of per-base mutation rates was tested for a range of population sizes, with the simulation being repeated and run 2000 times for each combination. The mutation rate by which 95% of the runs had lost each peak was recorded as a critical mutation rate, where a peak was considered to have not ever been lost only if there were individuals remaining on it at the end of the 10,000 generations. The dominance parameter λ was set to equal just below 1.0 as when $\lambda<1.0$, the fitness of both an individual's constituent sequences is taken into account to calculate an overall fitness score. The fitness of each individual was evaluated based on its constituent maternal and paternal sequences as per section 4.2.2.4. The method was run with sequence lengths in the range found in biology, as reported in chapter 7.

7.3 Results

The results were analysed by focusing on specific sequence lengths for which there was a comparable entry in both Table 7.1 and Table 7.2. This enabled plotting of

the critical mutation rate and error threshold according to the definitions presented in Chapters 5 and 6, along with a marker to represent an example mutation rate observed in nature for a gene of a similar length. Figures 7.1, 7.2, 7.3 and 7.4 each show the mutation rate at which 95% of the runs had lost either peak 0 (the critical mutation rate) or peak 1 (the error threshold) for varying population sizes. Figure 7.5 shows the maximal critical mutation rate and maximal error threshold produced by the simulation for each sequence length, plotted with biological mutation rates for comparable gene lengths taken from Tables 7.1 and 7.2 respectively. The maximal critical mutation rate and error threshold were chosen as this represents the value at which each curve has levelled out, applicable to the range of population sizes normally expected for each species without threat of extinction. Note the log scale used for the mutation rate as this enables the difference between the curves and the biological mutation rates to be seen clearly. For each sequence length it can be seen that both the critical mutation rate and error threshold curves can be approximated by an exponential function, and that the biological mutation rates are always lower than both curves. The transition from survival-of-the-flattest to survival-of-the-fittest is shown in three dimensions for sequence lengths 2000 and 20000 in Figures 7.6 and 7.7 respectively; increasing sequence length by a factor of ten does not affect the relationship between population size and mutation rate at varying percentage peak loss.

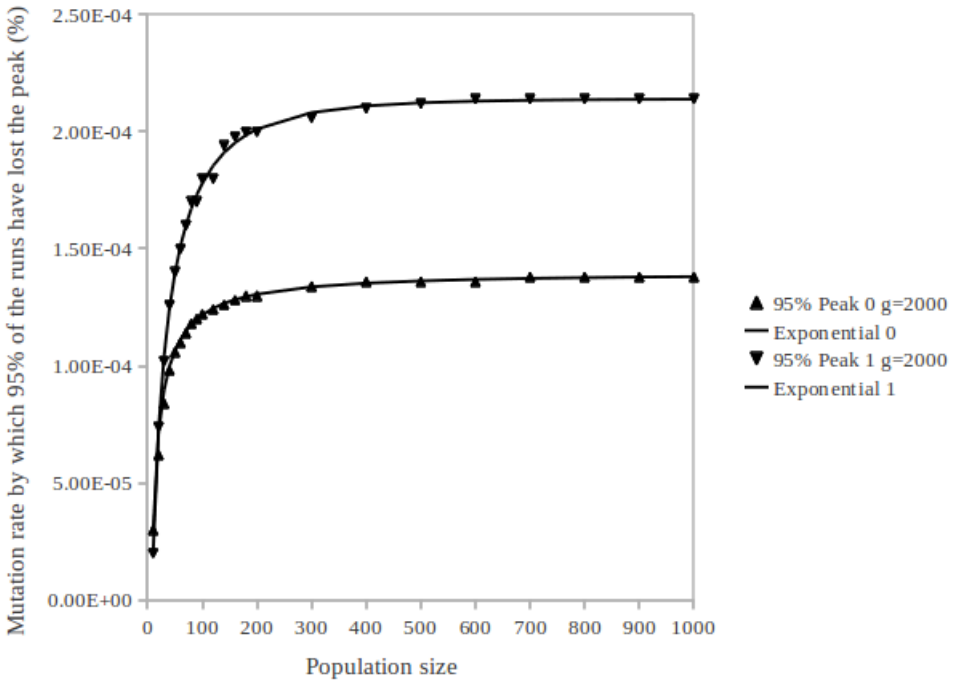

Figure 7.1: **Critical mutation rate and error threshold when the GA was run with a sequence length of 2000.** The exponential lines were obtained by curve-fitting using R with a least squares method (as per Figures 5.1 and 6.1).

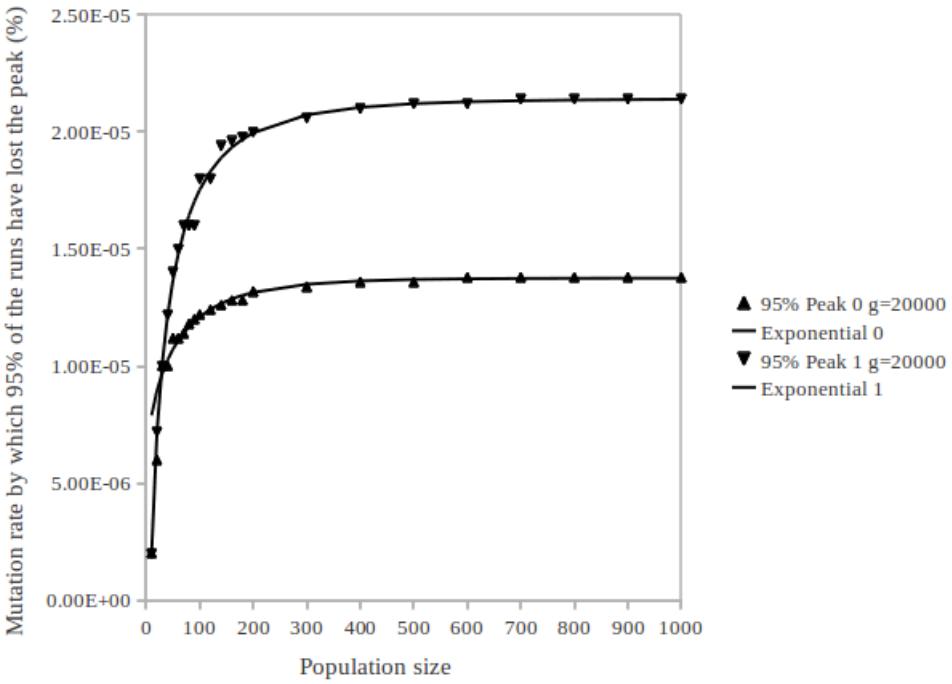

Figure 7.2: **Critical mutation rate and error threshold when the GA was run with a sequence length of 20000.** The exponential lines were obtained by curve-fitting using R with a least squares method (as per Figures 5.1 and 6.1).

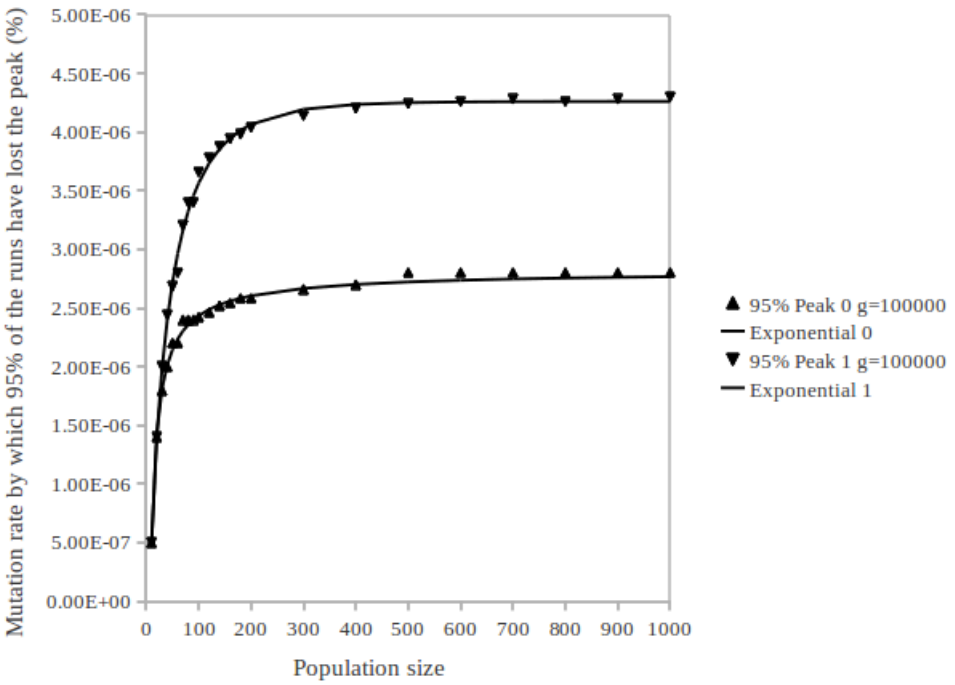

Figure 7.3: **Critical mutation rate and error threshold when the GA was run with a sequence length of 100000.** The exponential lines were obtained by curve-fitting using R with a least squares method (as per Figures 5.1 and 6.1).

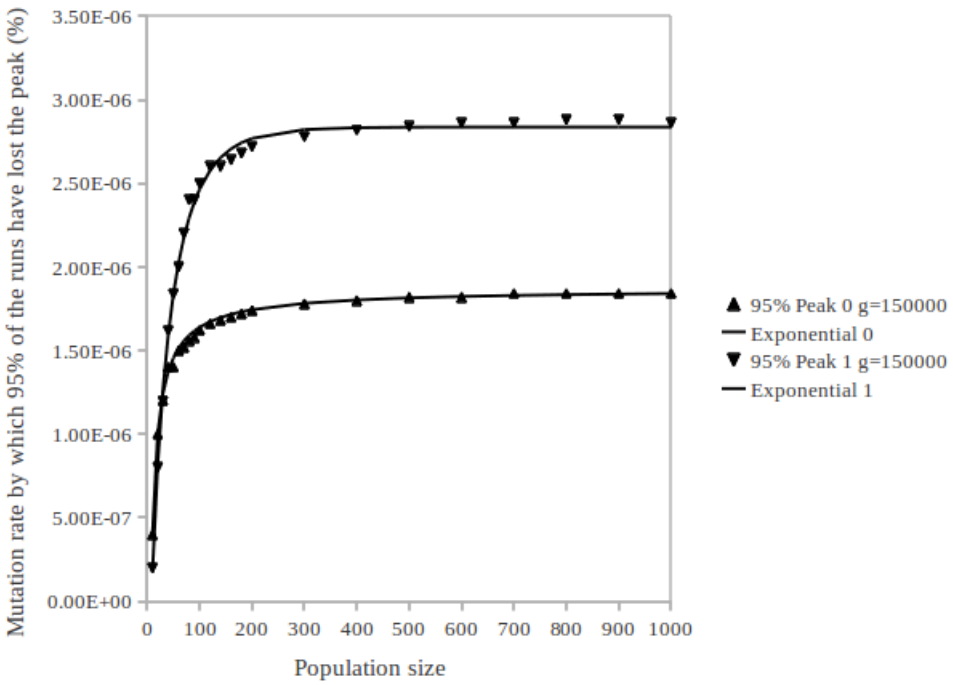

Figure 7.4: **Critical mutation rate and error threshold when the GA was run with a sequence length of 150000.** The exponential lines were obtained by curve-fitting using R with a least squares method (as per Figures 5.1 and 6.1).

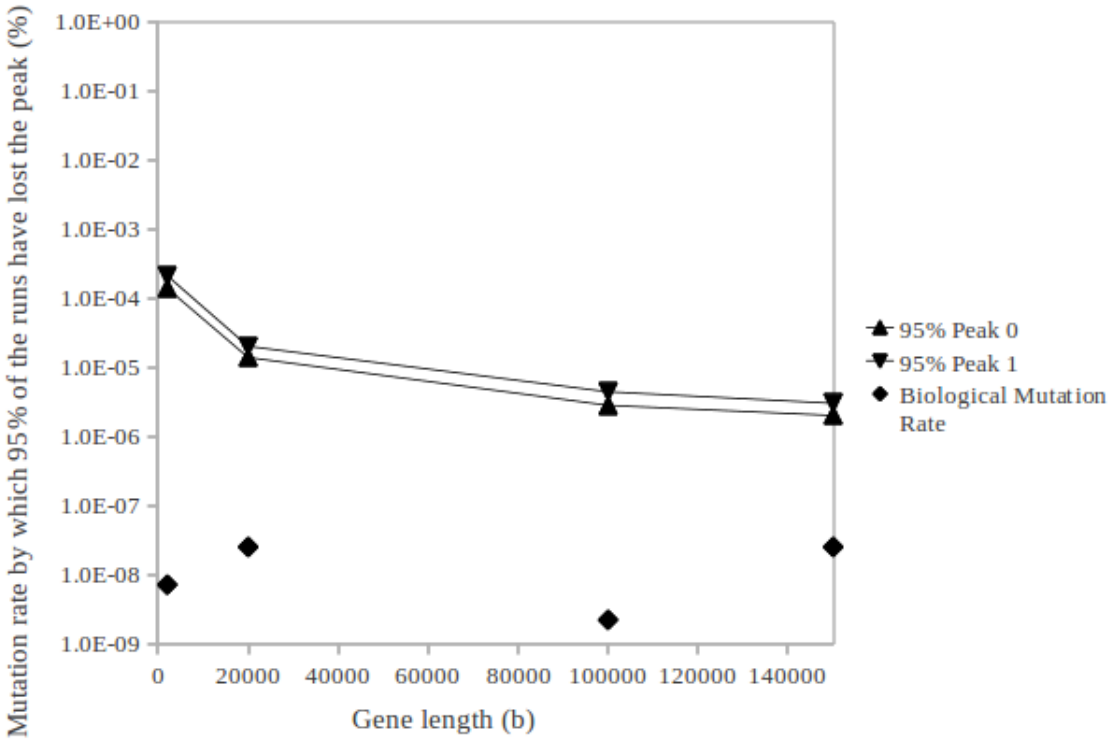

Figure 7.5: **Maximal critical mutation rate and error threshold plotted alongside biological mutation rates for varying sequence lengths.** Gene lengths close to those used in the simulation were selected from Table 7.2. The species associated with each gene length were then identified in Table 7.1, and the mutation rate plotted. Sequence length 2000 was matched with *Arabidopsis thaliana* which has a mean gene length of 2232 bp. The pre-selection, per base per generation mutation rate for *Arabidopsis thaliana* is 7.1×10^{-9}. Sequence length 20000 was matched with the average gene length of humans which is 27 kbp. The pre-selection, per base per generation mutation rate for humans is 2.5×10^{-8}. Sequence length 100000 was matched with the upper bound for the usual gene length range for flies and mammals which is 100 kb. The pre-selection, per base per genome per year mutation rate for the average mammal is 2.2×10^{-9}. Sequence length 150000 was matched with the largest human gene length in Table 7.2, the collagen family at 132.83 kbp. As before, the pre-selection, per base per generation mutation rate for humans is 2.5×10^{-8}.

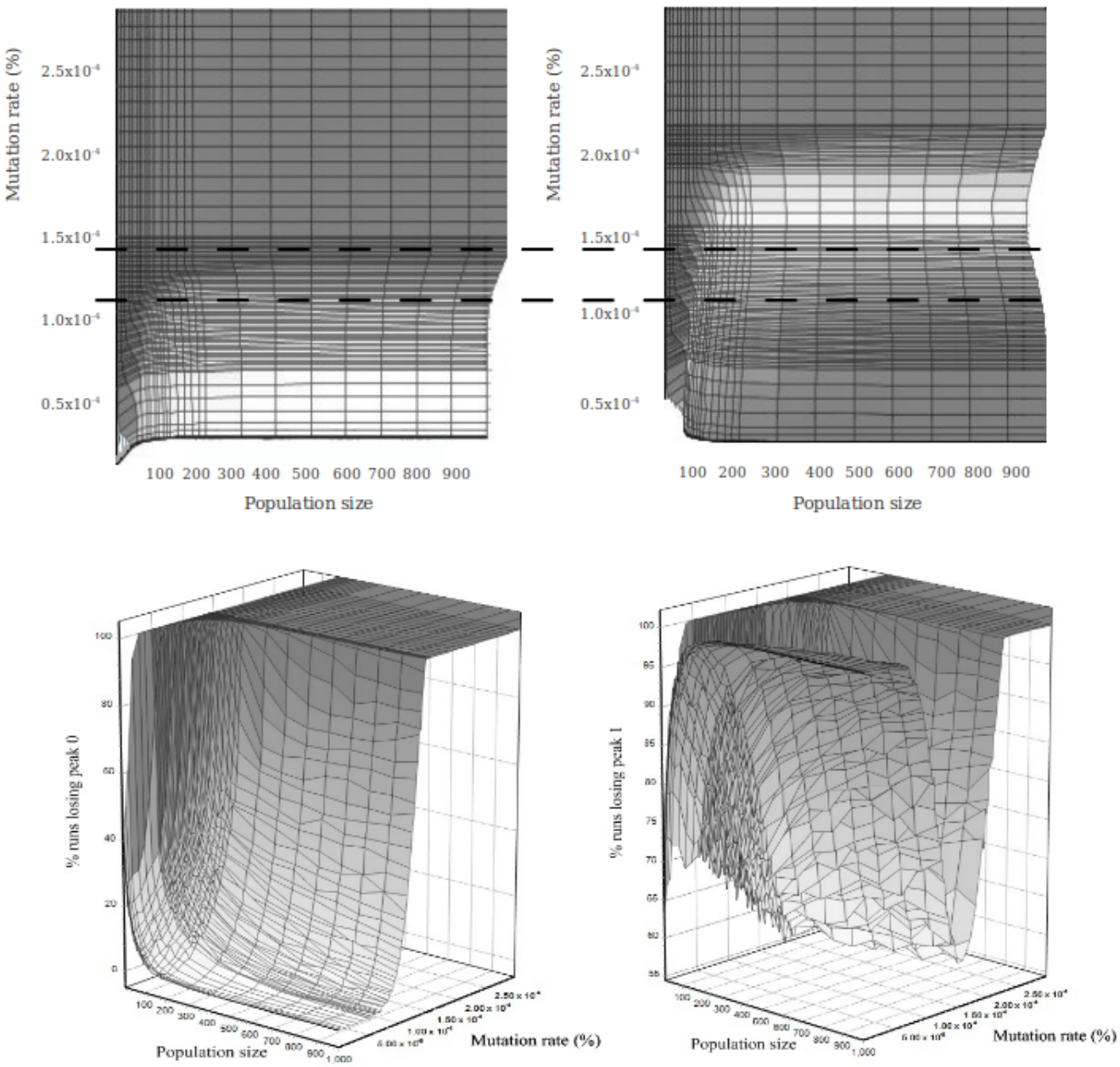

Figure 7.6: **Percentage of runs losing the peaks at different mutation rates and population sizes for sequence length 2000.** The results shown are for the diploid method with improved efficiency with λ just below 1 (see section 6.3.1 for an explanation as to why λ was not set to equal exactly 1), for peak 0 (a, left) and peak 1 (b, right). In the two upper projections the axis coming out of the page is the percentage of runs. The lower dashed line indicates the transition from survival-of-the-fittest to survival-of-the-flattest. The upper dashed line indicates the transition from survival-of-the-flattest to the error catastrophe. These graphs follow the same pattern as those in Figure 6.2.

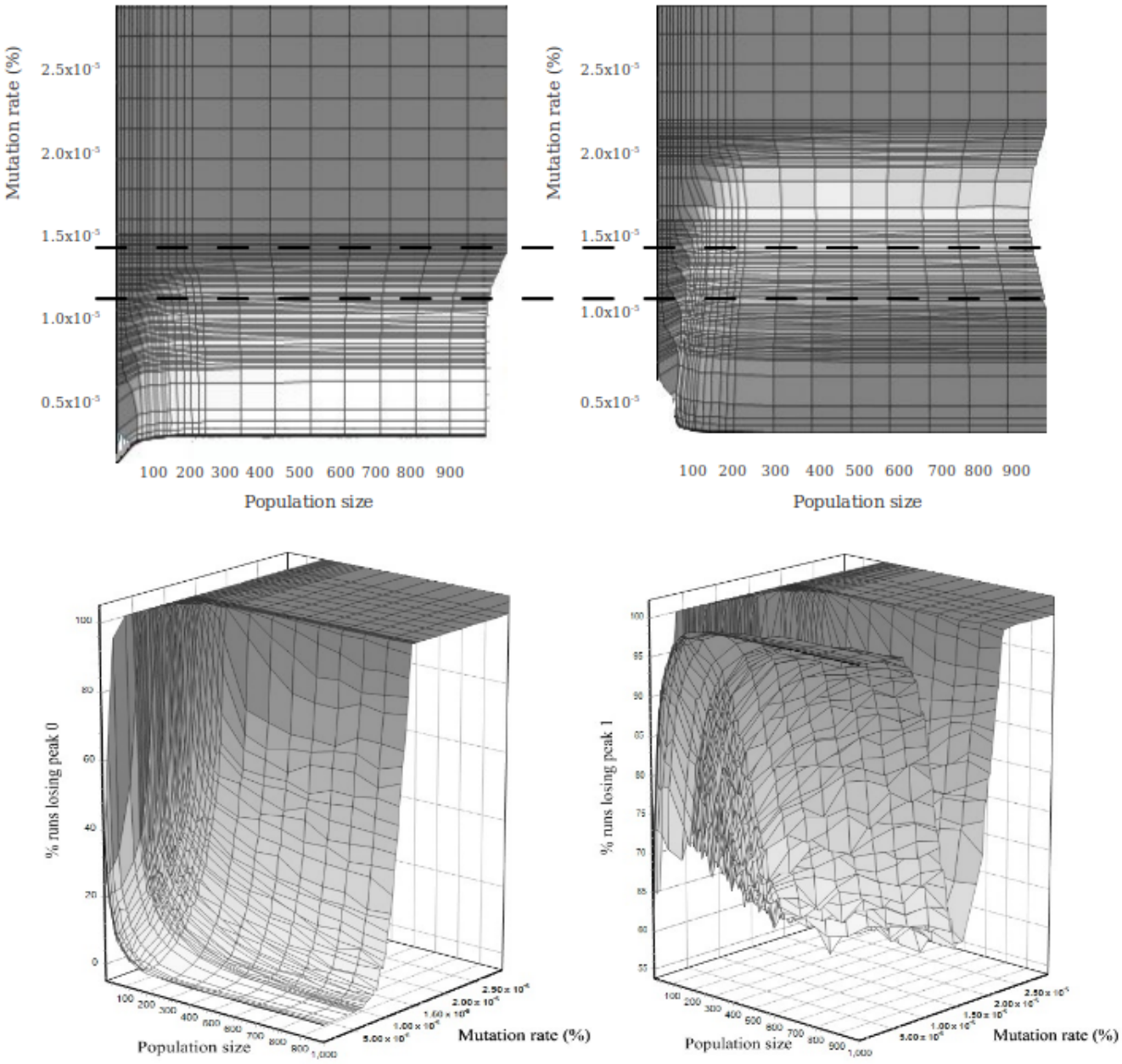

Figure 7.7: **Percentage of runs losing the peaks at different mutation rates and population sizes for sequence length 20000.** The results shown are for the diploid method with improved efficiency with λ just below 1 (see section 6.3.1 for an explanation as to why λ was not set to equal exactly 1), for peak 0 (a, left) and peak 1 (b, right). In the two upper projections the axis coming out of the page is the percentage of runs. The lower dashed line indicates the transition from survival-of-the-fittest to survival-of-the-flattest. The upper dashed line indicates the transition from survival-of-the-flattest to the error catastrophe. These graphs follow the same pattern as those in Figure 6.2 and Figure 7.6.

7.4 Discussion

Chapter 5 showed that population size influences the size of mutation rate that can be tolerated before fitter individuals are outcompeted by those that have a greater mutational robustness, despite the fact that previous studies had suggested this may not be the case [1]. The size of mutation rate at which the high, narrow peak and the lower, flatter peak were lost for increasing population sizes was shown to be approximated by an exponential function. The effect of population size on the size of mutation rate that could be tolerated before the population loses the fittest and the flattest peaks was shown to be particularly noticeable in small populations with 100 individuals or less. Chapter 6 extended the model to include diploidy, demonstrating that the critical mutation rate still has an exponential dependence on population size when using a genetic algorithm that is diploid and modelled closely on a real biological system. Chapter 7 has demonstrated that this model has relevance beyond that of artificial systems.

While many biological systems are more complex than the system presented here, the results shown in Figures 7.1, 7.2, 7.3, 7.4, and 7.5 show a link between the critical mutation rate and error threshold curves produced by the algorithmic method described in section 4.2.2.5, and the mutation rates observed in nature for species with comparable sequence lengths. Across each of the graphs it can be seen that increasing the sequence length lowers the magnitude of both the critical mutation rate and error threshold. The biological mutation rate associated with each sequence length is always below both the critical mutation rate and error threshold for each population size. A mutation rate above the error threshold would have indicated the population was experiencing an error catastrophe, while a mutation rate above the critical mutation rate but below the error threshold would have suggested the population was not evolving to maximal fitness; the observed biological mutation rates relate to established successful

populations therefore it was expected that the null hypothesis (that both the critical mutation rate and the error threshold would go below the typical mutation rates found in nature) could be rejected. More specifically, it can be seen from Figures 7.1, 7.2, 7.5, 7.6 and 7.7 that increasing the sequence length by a factor of 10 decreases the critical mutation rate and error threshold by an order of magnitude, meaning the null hypothesis (that increasing the sequence length would not lower both the critical mutation rate and error threshold in line with the exponential model) can also be rejected. A change in order of magnitude can also be seen in the observed values from nature. For example, in Table 7.2, the mean gene length of *Arabidopsis thaliana* is given as 2232 bp, while the average gene length of humans is just over 10 times longer at 27 kbp. In Table 7.1, the pre-selection, per base per generation mutation rate for *Arabidopsis thaliana* is given as 7.1×10^{-9}, while the pre selection, per base per generation mutation rate for humans is an order of magnitude higher at 2.5×10^{-8}. To give a further example, from Table 7.3, *Drosophila melanogaster* estimated genome size/gene number gives a sequence length of 8.82 kb. Human estimated genome size/gene number gives a sequence length approximately 10 times longer at 85.71 kb. Table 7.1 gives the pre selection, per base per generation mutation rate of *Drosophila melanogaster* as 4.65×10^{-9}, and that of humans an order of magnitude higher at 2.5×10^{-8}. This shows that a change in order of magnitude can also be seen when comparing like for like estimates of gene length and mutation rate found in nature, but further work will be required to establish whether a trend exists. It initially appears that while the critical mutation rate and error threshold do show a decreasing trend, the biological mutation rates do not; this may be due to the potential difference in distance between peaks for each gene. Varying this value according to the gene being modelled will be required to confirm this. The observation that biological mutation rates lie below the critical

mutation rate and error threshold produced by the algorithmic method suggests that the model may be applicable to natural evolving systems. Establishing the relationship between change in sequence length and magnitude of mutation rate, along with the effect of varying the distance between the peaks, will bring the model a step closer to bridging the gap between artificial and biological evolution.

7.5 Chapter Summary

Research question:

- Is the exponential model relevant to natural as well as artificial populations?

Novel results:

- Development of a faster algorithm capable of running experiments with parameter values within the range found in nature.
- Provided a link between the exponential model and mutation rates observed in biology. As expected, the biological mutation rates always lie below the critical mutation rate and error threshold.
- Demonstrated that increasing the sequence length by a factor of 10 decreases both the critical mutation rate and error threshold by an order of magnitude.

Chapters 5, 6 and 7 have presented the exponential model in terms of an artificial haploid algorithmic method, a diploid system modelled on the biological process of meiosis, and a diploid system using parameter values observed in nature. It has been shown that the critical mutation rate and error threshold curves produced by the diploid system using parameter values observed in nature are always higher than the equivalent biological mutation rates. This was expected, as optimal mutation rates for

any population will be within the range at which survival-of-the-fittest occurs. Chapter 8 considers the link between critical mutation rates and optimal mutation rate control theory.

Chapter 8

Critical Mutation Rates and Optimal Mutation Rate Control Theory

Optimal mutation rates have been studied extensively in the context of genetic algorithms and optimal genetic search [98, 97, 100]. Belavkin et al. [43] extend this theory to consider quasi-biological evolution *in silico*, and the potential to bridge the gap between the theory and natural organisms. It should be noted that Elizabeth Aston did not carry out the work published in [43], but was involved in some of the background reading and discussion and is therefore a contributing author. The discussion in section 8.3 of this chapter is distinct from [43] and was written solely by Elizabeth Aston.

To recap section 3.1.5, the most sensitive of parameters in a genetic algorithm (GA) is thought to be the mutation rate [96, 97]. It has been suggested that $1/L$ is a universal value for the per bit mutation rate in a GA, where L is sequence length [99, 97]. Mühlenbein [99] states that $\mu = 1/L$ is optimal for general unimodal functions (where a function $f(x)$ is unimodal if for a value y, it is monotonically increasing for $x \leq y$, and monotonically decreasing for $x \geq y$) [97]. Ochoa [97] found that a mutation rate of $1/L$ will produce optimal or near optimal results in a GA. They also found that

increasing the selection pressure increases the magnitude of optimal mutation rates, with some decrease in optimal mutation rate at small population sizes, concluding that a rate of $1/L$ will only be sub-optimal when the selection pressure is either extremely weak or extremely strong, or when the population size is very small [97]. Cervantes and Stephens [100] suggest that both the $1/L$ and error threshold heuristics are too high in landscapes with multiple peaks.

Belavkin et al. [43] used an *in silico* system of asexual self-replicating organisms with a monotonic fitness landscape to derive the probability of adaptation as a function of mutation rate. Theoretical results were produced and evaluated by using a Meta-GA to evolve optimal mutation rate functions, the details of which can be found in [43], and described below in section 8.2. The error threshold is dependent on the fitness landscape and has been shown to be related to the optimal mutation rate [80, 100]. Ochoa et al. [80] empirically demonstrated the relationship between optimal mutation rate and error threshold by independently identifying both values *in silico*, and then comparing them to each other. Clune et al. [101] used computer simulations to show that optimal mutation rates evolve when the landscape is smooth, but in the case of rugged fitness landscapes with wide valleys, less-than-optimal mutation rates are favoured. The relationship between the critical mutation rate, error threshold and optimal mutation rate control is discussed below in section 8.3.

8.1 Adaptation in Hamming Space

In section 3.1.2, Fisher's Geometric Model was introduced in terms of an infinite Euclidean space. It can be generalized to a finite Hamming space. Let the set of all sequences of letters from a finite alphabet $\{1, ..., \alpha\}$ with length l be denoted by $H_\alpha^l := \{1, ..., \alpha\}^l$. There are α^l points. The space H_α^l uses the Hamming metric

$d_H(a,b) := |\{i : a_i \neq b_i\}|$, which counts the number of different letters between sequences a and b (the Hamming distance) (Figure 8.1). In Figure 8.1, sequence a which is within sphere (T, n), i.e., $a \in S(\mathrm{T}, n)$, mutates into sequence b which is within sphere (T, m), i.e., $b \in S(\mathrm{T}, m)$, by distance $r = d_H(a,b)$.

Fisher's conclusion for Euclidean space as described in section 3.1.2 suggests that the probability of adaptation decreases with respect to r, i.e., as the magnitude of the change increases, the chance of improvement decreases until it reaches zero (or at least negligible). However, in Hamming space, the situation can reverse so that the probability of adaptation increases with r. This is due to the finiteness of Hamming space. Specifically, it is due to the fact that in a finite space, the interior of the sphere $S(\mathrm{T}, n)$ can be larger than its exterior. In an infinite space the exterior must always be larger than the interior as the exterior represents an infinite number of possible sequences; in a finite space there will reach a point where the number of possible sequences beyond the current point is outnumbered by the possible sequences between the current point and the optimum. It should be noted that while the Hamming space considers the difference between sequences of equal length, an alternative is the Levenshtein metric, which compares variable length sequences and counts the least number of substitutions, insertions and deletions.

8.2 Optimal Mutation Rate Control

Belavkin et al. [43] used an *in silico* system of asexual self-replicating organisms with a monotonic fitness landscape to derive the probability of adaptation as a function of mutation rate. The probability of adaptation is dependent on the rate of mutation. This introduces the possibility that organisms may maximize the expected fitness of their offspring through mutation rate control. This can be simulated *in silico* using a

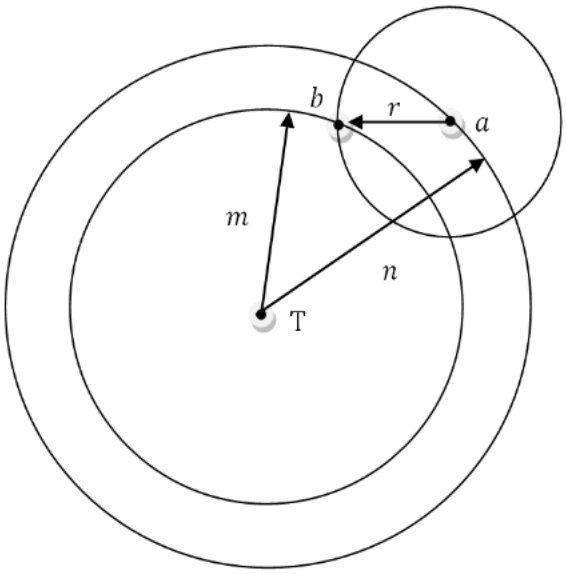

Figure 8.1: **Fisher's Geometric Model of Adaptation can be generalized to a Hamming space.** Here sequence $a \in S(\mathrm{T}, n)$ mutates into sequence $b \in S(\mathrm{T}, m)$ by $r = d_H(a, b)$.

system consisting of two genetic algorithms, known as a Meta-GA [43]. The Inner-GA is generational and uses no selection or recombination. In Belavkin et al. [43], there are 100 individuals, each of which is a sequence $\omega \in H_\alpha^l$. The initial population has the same number of individuals at each fitness value. Each individual is evolved for 500 generations by the process of point mutation, according to a mutation rate control function given by the Meta-GA. The Meta-GA is generational and uses tournament selection. There are 100 individuals, each of which is a mutation rate function $\mu(y)$ of fitness values y; the individuals are sequences of real values $\mu \in [0, 1]$ representing probabilities of mutation at different fitnesses. At each generation of the Meta-GA, 20 runs of the Inner-GA are evolved for 500 generations, with each run using a different mutation rate function $\mu(y)$ taken from the Meta-GA population. The Meta-GA then selects three individuals from its population of mutation rate functions at random and orders them according to fitness; the Inner-GA associated with each mutation rate function has its mean fitness calculated every generation, with each of its individuals having a fitness value determined by an arbitrary function $y = f(\omega)$. The least fit of the three mutation rate functions is replaced by a mutated crossover of the other two. Crossover combines the other two sequences, while mutation is done by adding a uniform random number to a randomly selected mutation rate on the mutation rate function (with lower and upper bounds of 0 and 1). This process repeats until every individual in the Meta-GA population has been selected, or until fewer than three individuals remain. The Meta-GA then returns the fittest of the mutation rate functions $\mu(y)$.

The mutation rate function evolved by the Meta-GA is dependent on the fitness landscape used in the Inner-GA. If fitness $f(\omega)$ corresponds to negative Hamming distance to the optimum $-d_H(\mathrm{T}, \omega)$, then the optimal mutation rate can be seen to

increase with $n = d_H(\mathrm{T}, \omega)$. This also applies to the population of mutation rate functions in the Meta-GA in the same type of landscape, as shown in Figure 8.2. Biologically relevant landscapes are likely to be more complex than the simple case where fitness equates to $-d_H(\mathrm{T}, \omega)$. In more rugged landscapes, if fitness $y = f(\omega)$ does not equate to $-d_H(\mathrm{T}, \omega)$, then fitness values of each sequence do not give complete information about the position of the sequence in the landscape. The fitness landscape defines a joint distribution of all the fitness values $y = f(\omega)$ and distances $n = d_H(\mathrm{T}, \omega)$ from the nearest optimum. Consider mutation of sequence a to sequence b at corresponding distances of $n = d_H(\mathrm{T}, a)$ and $m = d_H(\mathrm{T}, b)$ from the nearest optimum. Let their current fitness values be $y_t = f(a)$ and $y_{t+1} = f(b)$ respectively. The fitness and distance values for sequence b can be seen to be independent of those of its parent sequence a. The similarity of the transition probability between y_t and y_{t+1} and the transition probability between spheres of radius m and n from the optimum increases as sequences evolve closer to the optimum; optimal mutation rate control based on current fitness values should therefore resemble optimal mutation rate control based entirely on distance within some neighbourhood of the optimum.

8.2.1 Evolving Mutation Rate Control Functions in Biologically Relevant Landscapes

In biological systems, mutation is to some degree controlled by the organism [148, 43]. There is genetic variation in mutation rates within species and in the way mutation rates vary in a changing environment [43]. Different species have different rates of mutation (as discussed in chapter 7, Table 7.1), meaning mutation rate variation exists widely in nature. The existence of optimal mutation rates that are dependent on an individual's fitness implies that there may exist some level of mutation rate control

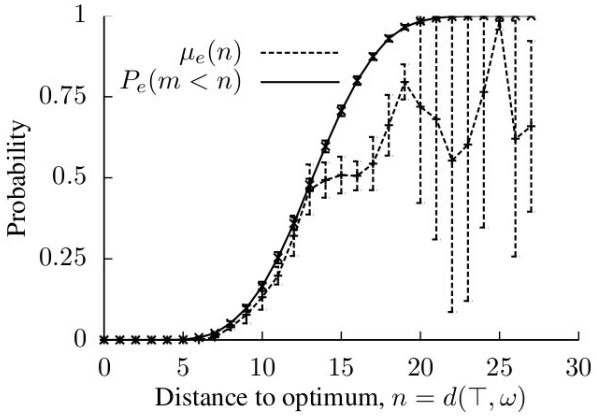

Figure 8.2: **Average of evolved mutation functions $\mu_e(n)$ and CDF $P_e(m < n)$ for fitness $-d_H(\mathbf{T}, \omega)$ in H_2^{30}.** Taken from [43]. The right hand side of the graph represents areas far away from the optimum that are less explored by the population of 100 individuals.

within biological organisms themselves [43]. Belavkin et al. [43] did multiple runs collecting multiple versions of evolved mutation control functions $\mu_e(x)$ and cumulative distribution functions $p_e(x_r > x)$ of observed fitness values for an aptamer landscape. An aptamer is a nucleic acid that can be selected to bind a particular target molecule [149]. Figure 8.3 shows the average of evolved mutation functions for the aptamer landscape H_4^{10}. This complete DNA-protein affinity landscape was described by Rowe et al. [150], and represents a rugged landscape with many local optima. Fitness is defined by the aptamer landscape rather than $f(\omega) = -d(\mathrm{T}, \omega)$. Figure 8.3 shows the CDF $p_e(x_r > x)$ to be approximately equal to the Meta-GA evolved function $\mu_e(x)$, following a comparable trend to that shown in Figure 8.2 for H_2^{30}. This provides evidence that the results discussed in section 8.2 have relevance to biology.

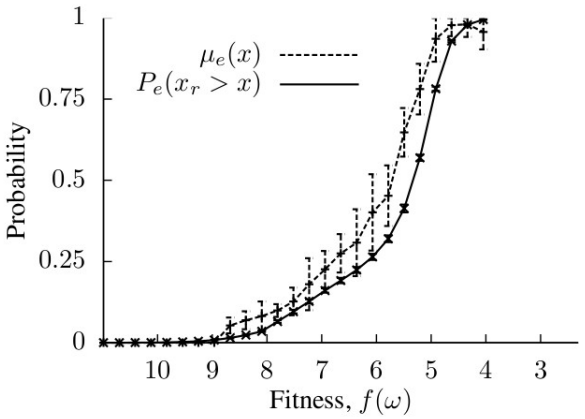

Figure 8.3: **Average of evolved mutation functions $\mu_e(x)$ and CDF $p_e(x_r > x)$ for fitness $f(\omega) = x$.** x is a random variable, and x_r are the observed fitness values. Results shown are for the aptamer 10 landscape [150] in H_4^{10} with 100 individuals.

8.3 Relating the Critical Mutation Rate

As the optimal mutation rate is related to the error threshold (section 3.1.5), and the critical mutation rate resembles an error threshold-like transition, one might reasonably expect some degree of relationship to exist between the error threshold and critical mutation rate. To examine this, consider the mutation rates obtained for varying haploid and diploid population sizes in chapter 5 and chapter 6. Figure 5.1 shows an exponential dependence on haploid population size for both the critical mutation rate and error threshold, with consistency observed for diploid populations in Figure 6.1. However, the difference between the absolute values of the respective critical mutation rate and error threshold curves is not consistent. This is particularly obvious in Figure 6.1, where the relationship between the critical mutation rate and error threshold curves is not consistent across all values for the dominance parameter. This means that, given an error threshold curve for a two-peak landscape such as that shown in

Figure 4.1, we can infer the shape of the respective critical mutation rate curve but not the magnitude. This implies some degree of relationship between the critical mutation rate and error threshold and their dependence on population size. Work by Eigen and Schuster [10] showed that viruses live very close to the error threshold. They are known to be efficient at evolving in new environments, implying a close relationship between optimal mutation rate and error threshold [80]. Given this, it should be possible to predict an approximate range for the error threshold based on known optimal mutation rates, and vice versa. As the magnitude of the error threshold says nothing about the magnitude of the critical mutation rate, the optimal mutation rate may not provide enough information to reasonably predict the critical mutation rate, and vice versa.

Given that the critical mutation rate occurs at the transition from survival-of-the-fittest to survival-of-the-flattest, it may be beneficial to think of the existence of an optimal mutation rate for each of these respective states. Consequently, the critical mutation rate may be considered to be analogous to the error threshold for the survival-of-the-fittest state. It seems reasonable to expect the relationship between the critical mutation rate and the survival-of-the-fittest optimal mutation rate to resemble that between the error threshold and the survival-of-the-flattest optimal mutation rate. While it will require further work to produce directly comparable data for all of these parameters, examining the relationship between the curves obtained for the critical mutation rate and optimal mutation rate suggests it may be possible to connect these critical parameters. Further work to tie together the system used to produce the optimal mutation rate curve and the algorithmic method used to produce the critical mutation rate curve would provide results that could potentially be directly joined to produce a 3D representation of the link between the optimal mutation rate, critical mutation rate and error threshold. This will require running the Meta-GA for different

population sizes as there is currently only data for population size 100 (as presented in Figures 8.2 and 8.3).

It has been established that in biological systems, mutation is to some degree controlled by the organism, and that the error threshold is dependent on the fitness landscape. The optimal mutation rate will also be dependent on the fitness landscape in that it is the mutation rate at which the population are best able to climb and maintain the peaks. This relationship is unidirectional as the fitness landscape does not vary according to the optimal mutation rate. The existence (and number) of critical mutation rates is also dependent on the fitness landscape. This provides a paradox. If the error threshold and critical mutation rate are dependent on the fitness landscape, and the fitness landscape is not subject to control by the organism, then this implies that the error threshold and critical mutation rate are also not subject to control by the organism. However, mutation is known to be controlled by the organism to some degree; if there is optimal mutation rate control, and there is a relationship between critical mutation rate, error threshold and optimal mutation rate, then it is reasonable to expect some degree of direct or indirect control of all of these parameters.

Chapter 9

Conclusions

9.1 Summary

In a fitness landscape, the fittest sequences are the 'peaks', while the lower fitness sequences occupy the 'valleys'. Sequence space is explored through evolution by mutation, recombination, selection and genetic drift in accordance to the fitness landscape. Mutation introduces variation, while selection acts to increase the frequency of fitter sequences. The first contribution of this study was the development of an algorithmic method operating at the level of the individual, in which selection is independent of the precise shape of the underlying landscape. The second contribution was the verification of this method using equations from analytical models (Equations 5.1 and 5.2) to produce comparable curves (Figure 5.2). Nowak and Schuster [32] present an analytical expression for the population size dependence of the error threshold (Equation 5.1). Ochoa et al. [119, 80] include a reformulation of the Nowak and Schuster analytical expression (Equation 5.2), in which they make explicit the reduction in the error threshold when moving from infinite populations to those of size N [119]. The observed consistency with the analytical error threshold models provided verification

for the critical mutation rate results and algorithmic method as a whole.

The third contribution of this work was to show that, for a haploid population and a two-peak landscape, the mutation rates at which the high, narrow peak and the lower, flatter peak are lost for increasing population sizes (of individuals of length 30) can be approximated by an exponential function, rejecting the null hypothesis that critical mutation rate has no dependence on population size in haploid populations. The effect of population size was particularly noticeable in populations of 100 individuals or less. The curve obtained for the critical mutation rate could be seen to flatten out to a greater degree than the curve obtained for the error threshold (see the faint lines in Figure 5.1). This explains why previous studies of larger populations have concluded that there is no relationship between the critical mutation rate and population size (e.g., [1]).

Using a genetic algorithm based on the biological process of meiosis, the fourth contribution was to demonstrate that the exponential relationship is conserved when moving from haploidy to diploidy, but that the critical mutation rate curves observed for a diploid system are lower than those observed for a haploid system (Figure 6.1), rejecting the null hypothesis that critical mutation rate has no dependence on population size in diploid populations. It has been suggested that there is an interaction between mutation rates and mating strategies in nature [122]. Haploid systems use between-individual recombination while diploid systems use within-individual recombination. Recombination lowers the mutation rate at which the error threshold occurs [124]. Assortative, non-random mating, in which individuals of a similar phenotype mate more often than expected by chance, is able to overcome this shift toward lower error threshold magnitudes induced by recombination [122]. Conversely, dissortative mating, in which dissimilar individuals mate more often, reduces the magnitude of

the error threshold. In the haploid system, the simulation starts with the population clustered at the two peaks. As the simulation is run, the population tends towards one of the peaks assuming the mutation rate does not exceed the error threshold. Recombination therefore tends to occur between sequences with similar fitnesses, and mating can be considered to be assortative. In the diploid system, the simulation starts with the population clustered at the two peaks, with individuals either completely at either peak, or with one sequence at one peak and one at the other. As fitness is calculated as a single value based on the fitness of an individual's two constituent sequences (see section 4.2.2.4), an individual can have, for example, a high fitness value but consist of two sequences in completely different parts of the fitness landscape. There is therefore a chance that the individuals selected to mate could have very different genetic make-ups; the degree of dissortative mating exceeds that of the haploid system. The difference in mating systems used by haploids and diploids is a potential reason for the difference in the curves shown in Figure 6.1. The fifth contribution of this work was to show that critical mutation rates are lower for diploid populations than haploid populations because of the difference in recombination (Figure 6.3). The null hypothesis (that the magnitude of the critical mutation rate and error threshold will not change when moving from haploidy to diploidy) can be rejected.

The sixth contribution of this work was the development and improvement of the definition of the critical mutation rate following analysis of the results. Previous studies have defined the critical mutation rate to be the midpoint between the highest mutation rate at which there is survival-of-the-fittest, and the lowest mutation rate at which there is survival-of-the-flattest [33, 1]. However, the results clearly show that there is a transition from survival-of-the-fittest to survival-of-the-flattest and subsequently to the error catastrophe (Figure 5.4). Figure 5.4(a) shows that 95% of the runs had

lost peak 0 within the duration of the simulation when the per-base mutation rate was 1.08%; just 52% of these runs lost the lower, flatter peak (peak 1). At this point, the transition from survival-of-the-fittest to survival-of-the-flattest is essentially complete. This can be considered as a critical mutation rate. Figure 5.4(b) shows that 95% of the runs had lost peak 1 within the duration of the simulation when the per-base mutation rate was 1.85%; 100% of these runs lost peak 0. This demonstrates that the transition from survival-of-the-flattest to the error catastrophe is essentially complete, with the population having almost entirely lost the ability to localize to either peak at this mutation rate. Figure 6.2 shows these transitions occurring in a diploid population, and demonstrates a relationship between the critical mutation rate and the error threshold. The highest point at lower mutation rates in (b) appears to correspond to where the curve in (a) starts to ascend. Likewise, by the time the curve in (b) has descended to its lowest, the curve in (a) has reached its highest. This shows the transition of the population favouring peak 0 to favouring peak 1. The transition occurs around the critical mutation rate. At less than 50% loss of peak 0, individuals are still moving from peak 1 to peak 0. The critical mutation rate concerns the loss of individuals from peak 0 to peak 1, therefore the critical mutation rate should not be considered to be at a point where there is still a significant transition in the other direction (implying there is still a peak 0 advantage). In the top graph in Figure 6.2 (b), it can be seen that for smaller population sizes (less than 50), the curve does not fall below approximately 70% loss of peak 1. Considering the equivalent portion of the graph, Figure 6.2 (a) suggests that considering a peak loss of anything much less than 50% will be redundant when the population is small. The critical mutation rate should be considered not as a single value at the midpoint, but rather as *lying within a range of values with a lower limit of 50% loss of the high, narrow peak.*

The seventh contribution of this work was the demonstration of the relevance of the exponential model to natural systems, and the eigth the development of a faster algorithm capable of running experiments with parameter values within the range found in nature. The results shown in Figures 7.1, 7.2, 7.3, 7.4 and 7.5 show a link between the critical mutation rate and error threshold curves produced by the algorithmic method described in section 4.2.2.5, and the mutation rates observed in nature for species with comparable gene lengths. Across each of the graphs it can be seen that increasing the sequence length lowers the magnitude of both the critical mutation rate and error threshold, with the biological mutation rate associated with each sequence length always below both the maximal critical mutation rate and maximal error threshold. More specifically, it can be seen from Figures 7.1, 7.2, 7.5, 7.6 and 7.7 that increasing the sequence length by a factor of 10 decreases the critical mutation rate and error threshold by an order of magnitude, making up the ninth contribution of this work. This observation suggests that the model may be applicable to natural evolving systems. Table 7.2 shows the mean gene length of *Arabidopsis thaliana* to be 2232 bp, while the average gene length of humans is just over 10 times longer at 27 kbp. Table 7.1 shows the pre-selection, per base per generation mutation rate for *Arabidopsis thaliana* to be 7.1×10^{-9}, while the pre-selection, per base per generation mutation rate for humans is an order of magnitude higher at 2.5×10^{-8}. It initially appears that the biological mutation rates do not follow the trend observed for the critical mutation rate and error threshold. This may be due to a potential variation in the distance between peaks for each gene. Beginning the process of linking together these critical mutation rate and error threshold results with optimal mutation rate control theory is the tenth contribution of this work.

These contributions provide the key insight that the critical mutation rate, at which

individuals with greater robustness to mutation are favoured over individuals with greater fitness, has an exponential dependence on population size in both haploid and diploid populations, the latter in a system modelled on the biological process of meiosis, with parameter values within the range found in nature. This is in contrast to previous studies which identified that critical mutation rate was independent of population size. The results show the effect of population size to be particularly strong in small populations with 100 individuals or less. When a population's size drops to this level, its critical mutation rate can be exceeded (in the absence of rapid mutation rate control) leading to loss of genetic material and a feedback spiral into further population size decline, genetic loss and on toward extinction. Population decline can lead to loss of fit genetic material that may be difficult to recover in very small populations. This has not identified a threshold for extinction, but has highlighted the fact that smaller populations experience error catastrophes, during which a population shifts in genotype space to areas of the landscape with lower fitness, at lower mutation rates. Such shifts indicate a population has become less well adapted to the current environment; smaller populations are at greater risk of extinction due to the presence of fewer individuals in the first place, with a smaller gene pool. Future work may determine the effect this has on population extinction and recovery, using parameter values within ranges found in nature. Testing the efficacy of different population management and conservation strategies (such as combining or mixing multiple small populations) on populations of varying sizes could also highlight the importance of considering population size and its relationship to genetic loss, as demonstrated here, during the decision making process.

9.2 Contributions

This work has made ten contributions which together provide the key insight that the critical mutation rate, at which individuals with greater robustness to mutation are favoured over individuals with greater fitness, has an exponential dependence on population size in both haploid and diploid populations. Importantly, this has been shown to have relevance in both artificial and biological populations. Specifically, these contributions are as follows:

1. An algorithmic method, operating at the level of the individual, which does not rely on the precise details of the underlying fitness landscape and is therefore capable of providing widely applicable results.

2. Verification of the method against analytical models for the error threshold, providing confidence in our subsequent results.

3. The discovery of an exponential relationship between the critical mutation rate and population size in haploid populations. The results show the effect of population size to be particularly strong in small populations with 100 individuals or less.

4. The result that this is conserved when moving from haploidy to diploidy but that the critical mutation rate and error threshold are both unexpectedly lower in the latter case. This provides the key insight that the critical mutation rate, at which individuals with greater robustness to mutation are favoured over individuals with greater fitness, has an exponential dependence on population size in both haploid and diploid populations, the latter in a system modelled on the biological process of meiosis. This is in contrast to previous studies which identified that critical mutation rate was independent of population size.

5. Other studies have suggested different types of recombination affect the error threshold. Extending from this, it has been shown that critical mutation rates are lower for diploid populations than haploid populations because of the difference in recombination.

6. An analysis of the transition from critical mutation rate to error threshold (survival-of-the-fittest to survival-of-the-flattest) which provides for an improvement on previous definitions of the critical mutation rate.

7. Provided a link between the exponential model and mutation rates observed in nature. As expected, the biological mutation rates always lie below the critical mutation rate and error threshold.

8. Development of a faster algorithm capable of running experiments with parameter values within the range found in nature.

9. Demonstrated that increasing the sequence length by a factor of 10 decreases both the critical mutation rate and error threshold by an order of magnitude.

10. Begun to link together the critical mutation rate, error threshold, and optimal mutation rate.

9.3 Limitations

Evolution in biological populations is complex, therefore designing and developing artificial systems to model the process is a challenge. The work presented in this thesis has begun the process but has been subject to limitations. To begin such a challenging process, it was necessary to first start with the simplest system in which survival-of-the-flattest could occur. This meant use of a two-peak landscape. Biological landscapes

can be more complex than this, therefore it may be beneficial to generalize this work to include multiple peaks. It was also limited to cases where genes are independent of other genes. This does not account for epistasis, where the expression of one gene is dependent on the presence of one or more other genes. Again, this was due to using the simplest system in which survival-of-the-flattest could occur.

In terms of mutation, this work only considered point mutations in which one base is changed to another base. It did not take into account other types of mutation such as insertions or deletions. Much of the theory focuses on point mutations, and mutation rates are often given per base, but this may be considered a limitation. Further to this, all point mutations are considered equal. In biology, the location of a point mutation can influence the effect that mutation has on the fitness of the individual.

One of the aims of this work was to use parameter values from nature to verify that the simulation models used and the results obtained have relevance to biological systems. While this was achieved to a reasonable degree in chapter 7, one limitation was the range of biological data available. It was hoped there would be a wide range of case studies available to examine and directly link to the exponential model, but a review of the literature provided very few. Further review will be required as future work.

9.4 Future Work

This work focused on the simplest case in which survival-of-the-flattest can occur, with the simplest type of mutation, and with all sequences as independent entities. This was necessary to identify the critical mutation rate and establish its relationship to population size without introducing too many additional variables. However, now that this relationship is known, it may be beneficial to consider introducing additional factors to

the algorithmic method such as insertions and deletions, epistasis, and the notion that some areas of the sequence can be more affected by mutation that others. Generalization to multiple-peak landscapes and the subsequent identification of multiple critical mutation rates and their relationship would also be an obvious step forward from this work. All of these additions will bring this work another step closer to bridging the gap between computational and biological evolution.

In addition to the potential for extending the algorithmic method, some direct areas for investigation have been identified through examination of the results presented in this thesis. This includes further examination of the role of haploidy and diploidy in evolution; the difference in mating systems used by haploids and diploids is a potential reason for the difference in the curves shown in Figure 6.1, as demonstrated by the results in section 6.3.2. It would be useful to explore this relationship further by using parameter values from nature for species that have a haploid and diploid stage in their life cycle, e.g., the yeast *Saccharomyces cerevisiae*.

Biological evolutionary systems are complex. Genes get copied and mutated through the process of transcription and translation by means of interactions between genetic sequences and proteins. It is possible to model these underlying processes *in silico* as demonstrated by Jenkins and Stekel's [151] model of evolving transcription factor control mechanisms in prokaryotic cells. The diploid method described in section 4.2.2.3 was modelled on the biological process of meiosis. This could be developed further to incorporate modelling of the mechanisms involved in meiosis by introducing parameters such as binding affinities of the transcription factor proteins involved in copying the genetic material. On a larger scale, it would be beneficial to begin to introduce more complexity by incorporating factors into the diploid method such as epistasis, different amounts of neutrality, varying proportions of non-coding sequences, and mul-

tiple fitness peaks in the landscape. Manrubia and Cuesta [79] suggest that there is virtually no single trait in multicellular animals or plants which is not dependent on a combination of genes acting together. Incorporating epistasis may be done by use of the NK landscape, introduced by Kauffman and Weinberger [152]. N is gene length and K is the number of other genes that influence that gene's fitness (the number of connections) [40]. Increasing K makes the landscape go from smooth to rugged. It would also be beneficial to experiment with the distance between the two peaks in the two-peak landscape. Tables 7.4 and 7.5 show a range of potential distances between the peaks both less and greater than the 10 base difference used in Chapters 5, 6 and 7. It is possible this parameter may affect the magnitude of both the critical mutation rate and error threshold curves and explain why they were orders of magnitude above the mutation rates observed in biology presented in chapter 7. It is also a potential explanation for why increasing the gene length by a factor of 10 decreases both the critical mutation rate and error threshold produced by the algorithmic method by an order of magnitude, but the biological mutation rates do not initially appear to follow this trend. To establish if there is any trend, it will be necessary to run the simulation using the gene lengths listed in Tables 7.2 and 7.3, and the distances between peaks listed in Tables 7.4 and 7.5.

This work only considered point mutations in which one base is changed to another base. It did not take into account other types of mutation such as insertions or deletions. As discussed in section 9.3, much of the theory focuses on point mutations, and mutation rates are often given per base, but introducing additional types of mutation may be a useful further step towards bridging the gap between artificial and biological evolution. Further to this, this work considers all point mutations as equal whereas in biology, the location of a point mutation can influence the effect that mutation has

on the fitness of the individual. It may be beneficial to emulate this by making the position of a point mutation in the simulation model influence its effect on fitness. This could be done by introducing a mutational 'hotspot' at a given position in the sequence and adjusting fitness according to the proximity of the point mutation to this position. This would also account for the potential for different parts of the genome to evolve at different rates [153, 154].

Chapter 8 began to link together the notion of critical mutation rates, error thresholds and optimal mutation rates. The next step may be to produce and directly compare values for optimal mutation rate, critical mutation rate, and error threshold as an extension of the discussion started in section 8.3. Further work to tie together the system used to produce the optimal mutation rate curve and the algorithmic method used to produce the critical mutation rate curve would provide results that could potentially be directly joined to produce a 3D representation of the link between the optimal mutation rate, critical mutation rate and error threshold. This will require running the Meta-GA for different population sizes as there is currently only data for population size 100 (as presented in Figures 8.2 and 8.3). It will also involve working out a way to identify an optimal mutation rate in terms of the simulation model.

Three of the major contributions of this thesis are the discovery of an exponential relationship between the critical mutation rate and population size in haploid populations, the conservation of this relationship when moving to diploidy, and the confirmation that the model is relevant to real biological populations. The ultimate aim of this is to produce a model that can be used to test the efficacy of different population management and conservation strategies (such as combining or mixing multiple small populations) on populations of varying sizes. After incorporating the additional complexity described earlier in this section, such as epistasis and the inclusion of mu-

tational hotspots, the next step towards achieving this aim is to match the output of the diploid method to a case study. When the diploid method is capable of taking parameter values relating to a case study and outputting comparable results, this will be an important step towards producing a model capable of predicting the fate of a population based on its current size and observed mutation rate.

9.5 Final Words

Studying evolution *in vivo* can be near infeasible due to time constraints and cost of resources. There have been many advances made with *in vitro* studies of microorganisms, but the challenge lies in bringing these results into the domain of larger species evolving over much longer timescales. *In silico* systems offer a method of modelling evolution with a wide range of parameter values, limited costs, and minimal time limitations. The big challenge is developing *in silico* systems that reliably and accurately model natural evolving systems. This thesis has presented an algorithmic method capable of modelling the phenomenon of survival-of-the-flattest that has been observed in digital organisms, theoretically, in simulated RNA evolution, and in RNA viruses. The method was developed to model the biological process of meiosis, and used with parameter values found in nature as an attempt to begin bridging the gap between computational and biological evolution. In my opinion, methods such as this will become increasingly important as it becomes necessary to model species that cannot be easily studied, with limited amounts of resources. The results presented in this thesis have highlighted the importance of considering population size and its relationship to genetic loss when developing population management and conservation strategies. Further development of such an algorithmic method may in the future enable modelling of specific species, prediction of their likely fate, and identification of the best possible

steps toward conservation.

Glossary

- *Allele* – One member of a pair (or any of the series) of genes occupying a specific spot on a chromosome (a locus) that controls the same trait.
- *Allozygous* - When two homozygous alleles are not identical by descent, but are instead unrelated. See *Autozygous.*
- *Amino acid* - The building blocks of proteins. Amino acids are coded for by DNA in living organisms.
- *Aptamer* - A nucleic acid that can be selected to bind a particular target molecule.
- *Autozygous* - Alleles that are identical by descent and homozygous. See *Allozygous.*
- *Base* - The building blocks of nucleic acids. In DNA, which is double stranded, bases pair up and the resulting base pairs form the DNA double helix.
- *Biological fitness* – The fitness of a genotype measures its relative ability to reproduce itself, compared to other genotypes. Fitness shows to what extent a genotype is favored by natural selection. Fitness values are between 0 and 1. The fittest individual has a fitness of 1, and the fitness of the other members of the population can be expressed as $1 - s$, where s is the selection coefficient.

- *Chromosome* - An organized structure made up of genetic material and proteins.

- *Correlated fitness landscape* - A non-random landscape that is tunable, such as the *NK* landscape. N is gene length and K is the number of other genes that influence that gene's fitness (the number of connections). Increasing K makes the landscape go from smooth to rugged.

- *Critical mutation rate* - A mutation rate at which the population loses its ability to remain on fitter peaks, but retains its ability to remain on flatter peaks of lower fitness; individuals with greater robustness to mutation are favoured over individuals with greater fitness. In the context of this thesis, the critical mutation rate should be considered to lie within a range of values with a lower limit of 50% loss of the high, narrow peak in a two-peak landscape.

- *Directional selection* - Also called positive selection. Alleles that increase the fitness of an individual will tend to increase in frequency until they replace the ancestral allele and become fixed in the population.

- *DNA* - Deoxyribonucleic acid. Double stranded. Encodes the genetic instructions in all living organisms and some viruses.

- *Dominance parameter (in the genetic algorithm)* – Value between 0 and 1, where 1 means take overall individual fitness to be that of the sequence with the highest fitness score of the two, and 0 means take that of the sequence with the lowest fitness score.

- *Dominant/recessive* – An allele is dominant if the phenotype of the heterozygote looks like the homozygote of that allele; the other allele in the heterozygote is referred to as recessive.

- *Effective population size* - The number of individuals in a population that contribute offspring to the next generation. In the results presented in the thesis 'population size' can be considered as the effective population size as the systems used ensure all individuals reproduce.

- *Electrophoresis* - Separates proteins by applying an electric charge across a gel. This causes the proteins to move across the gel based on their size and charge, where larger proteins move more slowly than smaller proteins. This method can also be used to separate DNA or RNA.

- *Epistasis* - When expression of one gene is dependent on the presence of one or more other genes.

- *Error threshold* - The mutation rate above which there is an error catastrophe and the population delocalizes across sequence space. Often not a precise value, but rather a range of mutation rates.

- *Euclidean space* - A space with a finite number of dimensions, in which any point can be represented by a coordinate.

- *Fitness (absolute)* - A measure of biological fitness expressed as the total number of gene copies transmitted to the subsequent generation or the total number of surviving offspring that an individual produces during its lifetime.

- *Fitness landscape* - Used to visualize the relationship between sequences and their fitness. Fitness landscapes are sometimes considered to resemble mountain ranges, with the fittest sequences at the peaks.

- *Fitness (relative)* - A measure of biological fitness expressed as the ratio of the absolute fitness of an individual (or of a genotype or of a phenotype) and the

absolute fitness of a reference individual (or of genotype or of phenotype).

- *Fitness score* – In the genetic algorithm, fitness is assigned a score. The value of the score is arbitrary; it is the relative value that is important.

- *Fixation* - The process by which a population (or subpopulation) becomes made up entirely of one type of allele due to a combination of random genetic drift and selection.

- *Fixation index* - The amount of differentiation in the population. Fixation index represents the difference between genetic sequences (genetic polymorphisms), and is related to identity by descent in that it measures how related two individuals from a subpopulation are in relation to the total population.

- *Fourfold site* - A position of a codon (three bases of a nucleic acid) is said to be a fourfold degenerate site if any nucleotide at this position specifies the same amino acid. For example, the third position of the glycine codons (GGA, GGG, GGC, GGU) is a fourfold degenerate site, because all nucleotide substitutions at this site are synonymous; they do not change the amino acid.

- *Genetic drift* - See *Random genetic drift.*

- *Genome size* – The number of bases in a particular genome.

- *Genotype* - The genetic makeup of an individual.

- *Hamming distance* - The number of differences between two sequences, e.g., 010232 is Hamming distance 2 away from 0**3**022**2**.

- *Hamming space* - A space in which points are separated based on the number of differences between them (the Hamming distance). The number of dimensions is

equal to the length of the sequences in the Hamming space.

- *Heterogeneous* – Non-uniform, diverse.
- *Heterozygote* - An individual having two different alleles at a genetic locus.
- *Homozygote* – An individual having two copies of the same allele at a locus.
- *Identical by descent* - Alleles that are both descended from and identical to an ancestral allele.
- *Information threshold* - The amount of information that can be maintained in a system. Linked to the *Error threshold*.
- *Locus* - A position on a chromosome.
- *Muller's ratchet* - The process by which the genomes of an asexual population irreversibly accumulate deleterious mutations. Contribute to mutational meltdown.
- *Mutational meltdown* - Meltdown occurs when a deleterious mutation becomes fixed in a population leading to reduced fitness and therefore reduction in population size. Mutations become fixed more rapidly the fewer individuals there are in the population; each time fixation of a deleterious mutation leads to reduction in population size it becomes easier for further deleterious mutations to become fixed leading to a potential downward spiral towards extinction. See *Muller's ratchet*.
- *Natural selection* - Selection as it occurs in nature. Individuals with higher fitness by definition reproduce at a higher rate. This leads to an increase in the pro-

portion of individuals in the next population that have their genes. It should be noted that fitness in this sense refers to the definition given for *Fitness (absolute)*.

- *Neutral* – Neutral mutations do not affect the fitness of the individual.
- *Nonsynonymous* - Mutation of a nucleotide that alters the coded amino acid of the resulting protein.
- *Per base mutation rate* – The number of times a single base will mutate in a given timeframe (e.g., per generation, per cell division, per year). If the term 'per base' is used alone, it refers to the number of times a base mutates in one reproduction.
- *Phenotype* - An individual's observable traits.
- *Point mutation* - Mutation of a single base.
- *Population size* – The number of individuals present in the population at a given time.
- *Pre-selection mutation rate* – The neutral mutation rate is independent of selection, as neutral mutations do not affect fitness. A proxy for neutral mutation rate is the substitution rate at fourfold sites, base positions in coding DNA that do not affect protein sequence and so will be under less selective pressure than other sites, e.g., see Kumar and Subramanian [133]. The fourfold sites also offer the advantage of being easily alignable for comparison. Fourfold sites have traditionally been seen as essentially free of selective constraint [128], at least in mammals where effective population sizes are often low and where mutations with a small effect on fitness should be expected to behave as neutral. The synonymous muta-

tion rate can be considered as a good representation of the neutral, pre-selection mutation rate.

- *Pseudogene* - A gene that does not code for a protein or is never expressed.
- *Purifying selection* – Also called negative selection. The selective removal of alleles that are deleterious.
- *Quasispecies* - A well-defined distribution of mutants generated by a mutation-selection process.
- *Random genetic drift* - The changes in allele frequency that occur by chance.
- *Recombination (biology)* – An event, occurring by the crossing-over of chromosomes during meiosis, in which DNA is exchanged between a pair of chromosomes. Thus two genes that were previously unlinked, being on separate chromosomes, can become linked because of recombination; and vice versa: linked genes may become unlinked. Like mutation, recombination is an important source of new variation for natural selection to work upon. However, also like mutation, recombination places a genetic load upon the population.
- *Recombination (in the genetic algorithm)* – Occurs through the process of crossover between individual sequences in the haploid genetic algorithm. Occurs though the process of crossover within individuals in the diploid genetic algorithm, between the constituent maternal and paternal sequences.
- *Recessive* - Refer to definition of 'dominant'.
- *Robustness* - Defined in terms of the average effect of a specified perturbation (such as a mutation) on the fitness of a specified genotype. The greater the robustness, the smaller the change in fitness.

- *RNA* - Ribonucleic acid. Single stranded. Encodes the genetic instructions in some viruses. In living organisms, DNA is translated to RNA before being translated to protein.

- *Selection* - The process by which sequences with higher fitness increase in proportion in a population. See also *Directional selection, Natural selection, Purifying selection* and *Selection pressure.*

- *Selection pressure* - The change in fitness caused by environmental conditions, e.g., competition with other individuals, sexual selection. The combined effect of multiple selection pressures determines the overall fitness of an individual. See *Selection.*

- *Sequence space* - When sequences are positioned so that immediate neighbours only have one bit difference.

- *Standing genetic variation* - the presence of more than one allele at a locus in a population.

- *Subpopulation* - A group of individuals typically distinct from the rest of the population geographically, with little migration occurring between different subpopulations.

- *Substitution rate* - The complete replacement of one allele by another within a population or species over evolutionary time.

- *Superiority* - The superiority parameter is usually denoted σ. This is generally calculated as the ratio of the fitness of the highest peak in the lanscape to the average fitness of all of the other peaks. In a single-peak landscape, this will

simply be the replication rate of the wild-type. In a two-peak landscape, it will be the ratio between the two peaks.

- *Synonymous/silent* - Substitions of one base for another are referred to as 'synonymous' or 'silent' if they occur in the exon of a gene and have no effect on the protein produced. Such mutations that occur in noncoding regions of the DNA are referred to as 'silent'. Synonymous substitutions are often assumed to be neutral, although certain codons are translated more efficiently than others.

- *Uncorrelated fitness landscape* - A random, 'rugged' landscape.

References

[1] I. Comas, A. Moya, and F. González-Candelas. Validating viral quasispecies with digital organisms: a re-examination of the critical mutation rate. *BMC Evolutionary Biology*, 5:5, 2005.

[2] C. Reidys, C.V Forst, and P. Schuster. Replication and mutation on neutral networks. *Bulletin of Mathematical Biology*, 63:57–94, 2001.

[3] M. Cargill, D. Altshuler, J. Ireland, P. Sklar, K. Ardlie, N. Patil, C. R. Lane, E. P. Lim, N. Kalyanaraman, J. Nemesh, L. Ziaugra, L. Friedland, A. Rolfe, J. Warrington, R. Lipshutz, G. Q. Daley, and E. S. Lander. Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nature Genetics*, 22:231–238, 1999.

[4] Birdlife International 2013. *Petroica traversi.* In *IUCN 2013. IUCN Red List of Threatened Species. Version 2013.2.*, 2013.

[5] S. Durant, L. Marker, N. Purchase, F. Belbachir, L. Hunter, C. Packer, C. Breitenmoser-Wursten, E. Sogbohossou, and H. Bauer. *Acinonyx jubatus.* In *IUCN 2013. IUCN Red List of Threatened Species. Version 2013.2.*, 2008.

[6] D. Ariano-Sánchez, J. Johnson, and M. Acevedo. *Abronia campbelli.* In *IUCN 2013. IUCN Red List of Threatened Species. Version 2013.2.*, 2013.

[7] A. Gonzalez, O. Ronce, R. Ferriere, and M.E. Hochberg. Evolutionary rescue: an emerging focus at the intersection between ecology and evolution. *Philosophical Transactions of the Royal Society B*, 368:20120404, 2013.

[8] M. Lynch, W. Sung, K. Morris, N. Coffey, C. R. Landry, E. B. Dopman, W. J. Dickinson, K. Okamoto, S. Kulkarni, D. L. Hartl, and W. K. Thomas. A genome-wide view of the spectrum of spontaneous mutations in yeast. *Proceedings of the National Academy of Sciences of the United States of America*, 105(27):9272–9277, 2008.

[9] S. Wright. The roles of mutation, inbreeding, crossbreeding, and selection in evolution. In *Proceedings of the Sixth International Congress on Genetics*, pages 355–366, 1932.

[10] M. Eigen and P. Schuster. *The hypercycle.* Springer, New York, 1979.

[11] E. Domingo and S. Wain-Hobson. The 30th anniversary of quasispecies. *EMBO Reports*, 10:444–448, 2009.

[12] M. Kimura and T. Maruyama. The mutational load with epistatic gene interactions in fitness. *Genetics*, 54:1337–1351, 1966.

[13] J. J. Bull, L. A. Meyers, and M. Lachmann. Quasispecies made simple. *PLoS Computational Biology*, 1(6):e61, 2005.

[14] M. A. Nowak. *Evolutionary dynamics: Exploring the equations of life.* Harvard University Press, 2006.

[15] M. A. Nowak. What is a quasispecies? *Trends in Ecology and Evolution*, 7:118–121, 1992.

[16] J. Masel and M. V. Trotter. Robustness and evolvability. *Trends in Genetics*, 26:406–414, 2010.

[17] M. Kimura. Evolutionary rate at the molecular level. *Nature*, 217:624–626, 1968.

[18] H. Allen Orr. The population genetics of beneficial mutations. *Philosophical Transactions of the Royal Society B*, 365:1195–1201, 2010.

[19] J. Jones and T. Soule. Comparing genetic robustness in generational vs. steady state evolutionary algorithms. In *Proceedings of the 8th annual conference on genetic and evolutionary computation*, pages 143–149, 2006.

[20] R.E. Lenski, J. E. Barrick, and C. Ofria. Balancing robustness and evolvability. *PLoS Biology*, 4:2190–2192, 2006.

[21] R. Sanjuán, J. M. Cuevas, V. Furió, E. C. Holmes, and A. Moya. Selection for robustness in mutagenized RNA viruses. *PLoS Genetics*, 3(6):e93, 2007.

[22] D. H. Reed. Relationship between population size and fitness. *Conservation Biology*, 19:563–568, 2005.

[23] D. E. Rozen, M. G. J. L. Habets, A. Handel, and J. A. G. M. de Visser. Heterogeneous adaptive trajectories of small populations on complex fitness landscapes. *PLoS One*, 3(3):e1715, 2008.

[24] R. Lande. Risk of population extinction from fixation of deleterious and reverse mutations. *Genetica*, 103:21–27, 1998.

[25] S. Wright. Evolution in mendelian populations. *Genetics*, 16:97–159, 1931.

[26] E. Tannenbaum and E. I. Shakhnovich. Solution of the quasispecies model for an arbitrary gene network. *Physical Review E*, 70:021903, 2004.

[27] N. Takeuchi and P. Hogeweg. Error-threshold exists in fitness landscapes with lethal mutants. *BMC Evolutionary Biology*, 7:15, 2007.

[28] P. Schuster. Genotypes and phenotypes in the evolution of molecules. *European Review*, 17(2):281–319, 2009.

[29] H. Tejero, A. Marin, and F. Montero. The relationship between error catastrophe, survival of the flattest, and natural selection. *BMC Evolutionary Biology*, 11:2, 2011.

[30] D. B. Saakian, E. Muñoz, C. Hu, and M. W. Deem. Quasispecies theory for multiple-peak fitness landscapes. *Physical Review E*, 73:041913, 2006.

[31] M. Eigen, J. McCaskill, and P. Schuster. Molecular quasispecies. *Journal of Physical Chemistry*, 92:6881–6891, 1988.

[32] M. A. Nowak and P. Schuster. Error thresholds of replication in finite populations: Mutation frequencies and the onset of Muller's ratchet. *Journal of Theoretical Biology*, 137:375–395, 1989.

[33] C. O. Wilke, J. L. Wang, C. Ofria, R. E. Lenski, and C. Adami. Evolution of digital organisms at high mutation rates leads to survival of the flattest. *Nature*, 412:331–333, 2001.

[34] C. O. Wilke. Quasispecies theory in the context of population genetics. *BMC Evolutionary Biology*, 5:44, 2005.

[35] J. Sardanyés, S. F. Elena, and R. V. Solé. Simple quasispecies models for the survival-of-the-flattest effect: The role of space. *Journal of Theoretical Biology*, 250:560–568, 2008.

[36] D. C. Krakauer and J.. B. Plotkin. Redundancy, antiredundancy, and the robustness of genomes. *Proceedings of the National Academy of Sciences of the United States of America*, 99(3):1405–1409, 2002.

[37] C. O. Wilke. Selection for fitness vs. selection for robustness in RNA secondary structure folding. *Evolution*, 55:2412–2420, 2001b.

[38] T. Wiehe, E. Baake, and P. Schuster. Error propagation in reproduction of diploid organisms: A case study on single peaked landscapes. *Journal of Theoretical Biology*, 177:1–15, 1995.

[39] J. H. Holland. Genetic algorithms. *Scientific American*, 267(1):66–72, 1992.

[40] G. Ochoa. Error thresholds in genetic algorithms. *Evolutionary Computation*, 14(2):157–182, 2006.

[41] A. Channon, E. Aston, C. Day, R. V. Belavkin, and C. G. Knight. Critical mutation rate has an exponential dependence on population size. In *Advances in Artificial Life, ECAL 2011: Proceedings of the Eleventh European Conference on the Synthesis and Simulation of Living Systems*, 2011.

[42] E. Aston, A. Channon, C. Day, and C. G. Knight. Critical mutation rate has an exponential dependence on population size in haploid and diploid populations. *PLoS ONE*, 8(12):e83438, 2013.

[43] R. V. Belavkin, A. Channon, E. Aston, J. Aston, and C. G. Knight. Theory and practice of optimal mutation rate control in Hamming spaces of DNA sequences. In *Advances in Artificial Life, ECAL 2011: Proceedings of the Eleventh European Conference on the Synthesis and Simulation of Living Systems*, 2011.

[44] C. Darwin and A. Wallace. On the tendency of species to form varieties; and on the perpetuation of varieties and species by natural means of selection. *Journal of the Proceedings of the Linnean Society of London. Zoology*, 3(9):45–62, 1858.

[45] R. A. Fisher. *The genetical theory of natural selection.* Oxford University Press, 1930.

[46] C. R. Darwin. *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life (3rd ed.).* London. John Murray, 1861.

[47] M. Ridley. *Evolution (2nd ed.).* Oxford University Press, 2004.

[48] S.J. Lolle, J.L. Victor, J.M. Young, and R.E. Pruitt. Genome-wide nonmendelian inheritance of extra-genomic information in *Arabidopsis*. *Nature*, 434:505–509, 2005.

[49] M. Rassoulzadegan, V. Grandjean, P. Gounon, S. Vincent, I. Gillot, and F. Cuzin. Rna-mediated non-mendelian inheritance of an epigenetic change in the mouse. *Nature*, 441:469–474, 2006.

[50] F. F. Costa. Non-coding rnas, epigenetics and complexity. *Gene*, 410:9–17, 2008.

[51] N. A. Campbell and J. B. Reece. *Biology.* Pearson Education, Inc., 7th edition, 2005.

[52] J. (ed.) Carroll. *On the origin of species.* Broadview Press, ON, Canada, 2003.

[53] K. M. Page and M. A. Nowak. Unifying evolutionary dynamics. *Journal of Theoretical Biology*, 219:93–98, 2002.

[54] D. L. Hartl and A. G. Clark. *Principles of population genetics*. Sinauer Associates, Inc., Sunderland, MA, 4th edition, 2007.

[55] R. Kliman, B. Sheehy, and J. Schultz. Genetic drift and effective population size. *Nature Education*, 1:3, 2008.

[56] A. Rosenberg and F. Bouchard. Fitness. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Fall 2010 edition, 2010.

[57] C. H. Waddington. Towards a theoretical biology. *Nature*, 218:525–527, 1968.

[58] D. C. Dennett. *Darwin's Dangerous Idea*. New York: Simon and Schuster, 1995.

[59] K. E. Boorman, B. E. Dodd, and B. E. Gilbey. A serum which demonstrates the co-dominance of the blood group gene O with A and B. *Annals of Human Genetics*, 14(1):201–208, 1947.

[60] R. A. Fisher. The possible modification of the response of the wild type to recurrent mutations. *American Naturalist*, 62:115–126, 1928.

[61] S. Billiard and V. Castric. Evidence for Fisher's dominance theory: how many 'special cases'? *Trends in Genetics*, 27(11):441–445, 2011.

[62] F. A. Kondrashov and E. V. Koonin. A common framework for understanding the origin of genetic dominance and evolutionary fates of gene duplications. *Trends in Genetics*, 20(7):287–291, 2004.

[63] H. Kacser and J. A. Burns. The control of flux. *Symposia of the Society for Experimental Biology*, 27:65–104, 1973.

[64] H. Kacser and J. A. Burns. The molecular basis of dominance. *Genetics*, 97:639–666, 1981.

[65] J. W. Porteous. A rational treatment of Mendelian genetics. *Theoretical Biology and Medical Modelling*, 1:6], 2004.

[66] J. B. S. Haldane. A mathematical theory of natural and artificial selection, part v: Selection and mutation. *Mathematical Proceedings of the Cambridge Philosophical Society*, 23:838–844, 1927.

[67] H. A. Orr. The distribution of fitness effects among beneficial mutations. *Genetics*, 163:1519–1526, 2003.

[68] J.H. Gillespie. Molecular evolution over the mutational landscape. *Evolution*, 38:1116–1129, 1984.

[69] R. Kassen and T. Bataillon. Distribution of fitness effects among beneficial mutations before selection in experimental populations of bacteria. *Nature Genetics*, 38:484–488, 2006.

[70] R. Sanjuán, A. Moya, and S.F. Elena. The distribution of fitness effects caused by single-nucleotide substitutions in an RNA virus. *Proceedings of the National Academy of Sciences of the United States of America*, 101:8396–8401, 2004.

[71] D. R. Rokyta, C. J. Beisel, P. Joyce, M. T. Ferris, C. L. Burch, and H. A. Wichman. Beneficial effects are not exponential for two viruses. *Journal of Molecular Evolution*, 67:368–376, 2008.

[72] G. R. Price. Fisher's 'fundamental theorem' made clear. *Annals of Human Genetics*, 36:129–140, 1972.

[73] H. A. Orr. The genetic theory of adaptation: a brief history. *Nature Reviews Genetics*, 6:119–127, 2005a.

[74] H. A. Orr. The distribution of fitness effects among beneficial mutations in Fisherś geometric model of adaptation. *Journal of Theoretical Biology*, 238:279–285, 2005b.

[75] H. A. Orr. Theories of adaptation: what they do and donť say. *Genetica*, 123:3–13, 2005c.

[76] J. Maynard Smith. Natural selection and the concept of a protein space. *Nature*, 225:563–564, 1970.

[77] M. Kimura. The neutral theory of molecular evolution: A review of recent evidence. *The Japanese Journal of Genetics*, 66(4):367–386, 1991.

[78] R. Shipman, M. Shackleton, , and I. Harvey. The use of neutral genotype-phenotype mappings for improved evolutionary search. *BT Technology Journal*, 18:103–111, 2000.

[79] S. C. Manrubia, , and J. A. Cuesta. Neutral networks of genotypes: Evolution behind the curtain. *Arbor*, 746:1051–1064, 2010.

[80] G. Ochoa, I. Harvey, and H. Buxton. Error thresholds and their relation to optimal mutation rates. In *Proceedings of the Fifth European Conference on Articial Life*, pages 54–63, 1999.

[81] P. Schuster, W. Fontana, P. F. Stadler, and I. Hofacker. From sequences to shapes and back: a case study in RNA secondary structures. *Proceedings of the Royal Society of London*, 255:279–284, 1994.

[82] P. Schuster and P. F. Stadler. Networks in molecular evolution: A common theme at all levels. *Complexity*, 8:34–42, 2003.

[83] P. Schuster. Landscapes and molecular evolution. *Physica D*, 107:351–365, 1997.

[84] A. Owen and I. Harvey. Adapting particle swarm optimisation for fitness landscapes with neutrality. In *Proceedings 2007 IEEE Swarm Intelligence Symposium*, pages 258–265, 2007.

[85] I. Harvey. Artificial evolution for real problems. In *Evolutionary Robotics: From Intelligent Robots to Artificial Life (ER'97)*, pages 127–149, 1997.

[86] L. Barnett. Netcrawling - optimal evolutionary search with neutral networks. In *In Proceedings of the 2001 Congress on Evolutionary Computation CEC2001*, pages 30–37, 2001.

[87] L. Duret. Neutral theory: The null hypothesis of molecular evolution. *Nature Education*, 1:1, 2008.

[88] J. Maynard Smith and E. Szathmáry. *The major transitions in evolution*. W. H. Freeman Spektrum, Oxford, 1995.

[89] N. Takeuchi, P. H. Poorthuis, and P. Hogeweg. Phenotypic error threshold; additivity and epistasis in rna evolution. *BMC Evolutionary Biology*, 5:9, 2005.

[90] J. Park, E. Muñoz, and M. W. Deem. Quasispecies theory for finite populations. *Physical Review E*, 81:011902, 2010.

[91] P. R. A. Campos and J. F. Fontanari. Finite-size scaling of the error threshold transition in finite populations. *Journal of Physics A: Mathematical and General*, 32:L1–L7, 1999.

[92] D. M. Lorenz, J-M. Park, and M. W. Deem. Evolutionary processes in finite populations. *Physical Review E*, 87:022704, 2013.

[93] P. A. P. Moran. *The Statistical Processes of Evolutionary Theory.* Clarendon Press, Oxford, 1962.

[94] C. C. Strelioff, R.E. Lenski, and C. Ofria. Evolutionary dynamics, epistatic interactions, and biological information. *Journal of Theoretical Biology*, 266:584–594, 2010.

[95] D.M. Weinreich and L. Chao. Rapid evolutionary escape by large populations from local fitness peaks is likely in nature. *Evolution*, 59:1175–1182, 2005.

[96] T. Bäck. *Evolutionary algorithms in theory and practice.* The Clarendon Press Oxford University Press, 1996.

[97] G. Ochoa. Setting the mutation rate: Scope and limitations of the 1/l heuristic. In *Proceedings of Genetic and Evolutionary Computation Conference (GECCO-2002)*, 2002.

[98] A. E. Eiben, J. McCaskill, and P. Schuster. Parameter control in evolutionary algorithms. *IEEE Transactions on Evolutionary Computation*, 3(2):124–141, 1999.

[99] H. Mühlenbein. How genetic algorithms really work: Mutation and hillclimbing. *Parallel Problem Solving from Nature*, 2:15–26, 1992.

[100] J. Cervantes and C. R. Stephens. "Optimal" mutation rates for genetic search. In *Proceedings of Genetic and Evolutionary Computation Conference (GECCO-2006)*, 2006.

[101] J. Clune, D. Misevic, C. Ofria, R. E. Lenski, S. F. Elena, and R. Sanjuán. Natural selection fails to optimize mutation rates for long-term adaptation on rugged fitness landscapes. *PLoS Computational Biology*, 4(9), 2008.

[102] T. Bäck and M. Schütz. Intelligent mutation rate control in canonical genetic algorithms. In *Proceedings of the Ninth International Symposiumon Foundations of Intelligent Systems*, volume 1079, 1996.

[103] G. Ochoa, I. Harvey, and H. Buxton. Optimal mutation rates and selection pressure in genetic algorithms. In *Proceedings of Genetic and Evolutionary Computation Conference (GECCO-2000)*, 2000.

[104] M. Kimura. *The neutral theory of molecular evolution.* Cambridge University Press, 1983.

[105] C.V. Forst, C. Reidys, and J. Weber. Evolutionary dynamics and optimization: Neutral networks as model-landscapes for RNA secondary-structure folding-landscapes. In *Advances in Artificial Life, vol. 929 of Lecture Notes in Artificial Intelligence*, 1995.

[106] M. A. Huynen, P. F. Stadler, and W. Fontana. Smoothness within ruggedness: The role of neutrality in adaptation. *Proceedings of the National Academy of Sciences of the United States of America*, 93:397–401, 1996.

[107] E. Bornberg-Bauer and L. Kramer. Robustness versus evolvability: a paradigm revisited. *HFSP Journal*, 4(3-4):105–108, 2010.

[108] C. O. Wilke. Adaptive evolution on neutral networks. *Bulletin of Mathematical Biology*, 63:715–730, 2001a.

[109] A.S. Lauring and R. Andino. Quasispecies theory and the behavior of RNA viruses. *PLoS Pathogens*, 6:7, 2010.

[110] E. van Nimwegan, J. P. Crutchfield, and M. Huynen. Neutral evolution of mutational robustness. *Proceedings of the National Academy of Sciences of the United States of America*, 96:9716–9720, 1999.

[111] J. D. Bloom, Z. Lu, D. Chen, A. Raval, O. S. Venturelli, and F. H. Arnold. Evolution favours protein mutational robustness in sufficiently large populations. *BMC Biology*, 5:29, 2007.

[112] D.A. Tallmon, G. Luikart, and R.S. Waples. The alluring simplicity and complex reality of genetic rescue. *Trends in Ecology and Evolution*, 19(9):489–496, 2004.

[113] J. J. Bull, R. Sanjuán, and C.O. Wilke. Theory of lethal mutagenesis for viruses. *Journal of Virology*, 81(6):2930–2939, 2007.

[114] G. Beslon, D. P. Parsons, Y. Sanchez-Dehesa, J.-M. Peña, and C. Knibbe. Scaling laws in bacterial genomes: A side-effect of selection of mutational robustness? *BioSystems*, 102:32–40, 2010.

[115] N. Metropolis and S. Ulam. The monte carlo method. *Journal of the American Statistical Association*, 44(247):335–341, 1949.

[116] T. S. Ray. Evolution, ecology and optimization of digital organisms. *Technical Report Working Paper*, pages 92–08–042, 1992.

[117] C. Adami and C. T. Brown. Evolutionary learning in the 2D artificial life system Avida. In *Proceedings of Artificial Life IV*, pages 377–381, 1994.

[118] T. S. Ray. *Artificial Life Programs and Evolution. In: Michael Ruse and Joseph Travis editors, Companion to Evolution.* Harvard University Press, 2009.

[119] G. Ochoa and I. Harvey. Recombination and error thresholds in finite populations. In *Foundations of Genetic Algorithms (FOGA-5)*, pages 245–264, 1998.

[120] B. Hutter, M. Bieg, V. Helms, and M. Paulsen. Divergence of imprinted genes during mammalian evolution. *BMC Evolutionary Biology*, 10:116, 2010.

[121] D. Alves and J. F. Fontanari. Error threshold in the evolution of diploid organisms. *Journal of Physics A: Mathematical and General*, 30:2601–2607, 1997.

[122] G. Ochoa and K. Jaffe. Assortative mating drastically alters the magnitude of error thresholds. In *Proceedings of the 9th international conference on Parallel Problem Solving from Nature*, pages 890–899, 2006.

[123] M. N. Jacobi and M. Nordahl. Quasispecies and recombination. *Theoretical Population Biology*, 70:479–485, 2006.

[124] M. C. Boerlijst, S. Bonhoeffer, and M. A. Nowak. Viral quasi-species and recombination. *Proceedings of the Royal Society of London B*, 263:1577–1584, 1996.

[125] R. Milo, P. Jorgenson, U. Moran, G. Weber, and M. Springer. Bionumbers: the database of key numbers in molecular and cell biology. *Nucleic Acids Research*, 38:D750–D753, 2010.

[126] C. G. Elsik, R. L. Tellam, and K. C. Worley. The genome sequence of taurine cattle: a window to ruminant biology and evolution. *Science*, 324:522–528, 2009.

[127] M.C. Whitlock, C.K. Griswold, and A.D. Peters. Compensating for the meltdown: The critical effective size of a population with deleterious and compensatory mutations. *Annales Zoologici Fennici*, 40:169–183, 2003.

[128] M. W. Nachman and S. L. Crowell. Estimate of the mutation rate per nucleotide in humans. *Genetics*, 156:297–304, 2000.

[129] C. F Baer, M. M. Miyamoto, and D. R. Denver. Mutation rate variation in multicellular eukaryotes: causes and consequences. *Nature Reviews Genetics*, 8:619, 2007.

[130] J. W. Drake, B. Charlesworth, D. Charlesworth, and J. F. Crow. Rates of spontaneous mutation. *Genetics*, 148:1667–1686, 1998.

[131] M. Lynch. Evolution of the mutation rate. *Trends in Genetics*, 26:345–352, 2010.

[132] Y. Xue, Q. Wang, Q. Long, B. L. Ng, H. Swerdlow, J. Burton, C. Skuce, R. Taylor, Z. Abdellah, Y. Zhao, Asan, D. G. MacArthur, M. A. Quail, N. P. Carter, H. Yang, and C Tyler-Smith. Human Y chromosome base-substitution mutation rate measured by direct sequencing in a deep-rooting pedigree. *Current Biology*, 19:1453–1457, 2009.

[133] S. Kumar and S. Subramanian. Mutation rates in mammalian genomes. *Proceedings of the National Academy of Sciences of the United States of America*, 99(2):803–808, 2002.

[134] P. D. Keightley, U. Trivedi, M. Thomson, F. Oliver, S. Kumar, and M. L. Blaxter. Analysis of the genome sequences of three drosophila melanogaster spontaneous mutation accumulation lines. *Genome Research*, 19:1195–1201, 2009.

[135] D. R. Denver, K. Morris, M. Lynch, and W. K. Thomas. High mutation rate and predominance of insertions in the caenorhabditis elegans nuclear genome. *Nature*, 430:679–682, 2004.

[136] C. Haag-Liautard, N. Coffey, D. Houle, M. Lynch, B. Charlesworth, and P. D. Keightley. Direct estimation of the mitochondrial dna mutation rate in drosophila melanogaster. *PLoS Biology*, 6(8):e204, 2008.

[137] S. Ossowski, K. Schneeberger, J. I. Lucas-Lledó, N. Warthmann, R. M. Clark, R. G. Shaw, D. Weigel, and M. Lynch. The rate and molecular spectrum of spontaneous mutations in arabidopsis thaliana. *Science*, 327:92–94, 2010.

[138] R. M. Durbin, D. Altshuler, G. R. Abecasis, D. R. Bentley, A. Chakravarti, A. G. Clark, and F. S. Collins. A map of human genome variation from population-scale sequencing. *Nature*, 467:1061–1073, 2010.

[139] M. Lynch. Rate, molecular spectrum, and consequences of human mutation. *Proceedings of the National Academy of Sciences of the United States of America*, 107(37):16013–16015, 2010.

[140] E. Derelle, C. Ferraz, S. Rombauts, P. Rouzé, A. Z. Worden, S. Robbens, F. Partensky, S. Degroeve, S. Echeynié, R. Cooke, Y. Saeys, J. Wuyts, K. Jabbari, C. Bowler, O. Panaud, B. Piégu, S. G. Ball, J. P. Ral, F. Y. Bouget, G. Piganeau, B. De Baets, A. Picard, M. Delseny, J. Demaille, Y. Van de Peer, and H. Moreau. Genome analysis of the smallest free-living eukaryote ostreococcus tauri unveils many unique features. *Proceedings of the National Academy of Sciences of the United States of America*, 103(31):11647–52, 2006.

[141] V. K. Sharma, S. K. Brahmachari, and S. Ramachandran. (tg/ca)n repeats in human gene families: abundance and selective patterns of distribution according to function and gene length. *BMC Genomics*, 6:83, 2005.

[142] B. Lewin. *Genes IX*. Jones and Bartlett Learning, 2008.

[143] L. Xu, H. Chen, X. Hu, R. Zhang, Z. Zhang, and Z. Luo. Average gene length is highly conserved in prokaryotes and eukaryotes and diverges only between the two kingdoms. *Molecular Biology and Evolution*, 23, 2006.

[144] M. M. Sachs, E. S. Dennis, and Peacock W. J. Gerlach, W. L. and. Two alleles of maize alcohol dehydrogenase 1 have 3' structural and poly(a) addition polymorphisms. *Genetics*, 113:449–467, 1986.

[145] G. T. Bryan, K. Wu, L. Farrall, Y. Jia, H. P. Hershey, S. A. McAdams, K. N. Faulk, G. K. Donaldson, R. Tarchini, and B. Valent. A single amino acid difference distinguishes resistant and susceptible alleles of the rice blast resistance gene pi-ta. *The Plant Cell*, 12:2033–2045, 2000.

[146] G. Ramkumar, A. K. Biswal, K. Madhan Mohan, K. Sakthivel, A. K. P. Sivaranjani, C. N. Neeraja, T. Ram, S. M. Balachandran, R. M. Sundaram, M. S. Prasad, B. C. Viraktamath, and M. S. Madhav. Identifying novel alleles of rice blast resistance genes pikh and pita through allele mining. *International Rice Research Notes*, 2010.

[147] P. M. Cummings and M. T. Clegg. Nucleotide sequence diversity at the alcohol dehydrogenase 1 locus in wild barley (Hordeum vulgare ssp. spontaneum): An evaluation of the background selection hypothesis. *Proceedings of the National Academy of Sciences of the United States of America*, 95:5637–5642, 1998.

[148] R. Hakem. DNA-damage repair; the good, the bad, and the ugly. *EMBO Journal*, 27(4), 2008.

[149] C.G. Knight, M. Platt, W. Rowe, D.C. Wedge, F. Khan, P.J. Day, A. McShea, J. Knowles, and D.B. Kell. Array-based evolution of DNA aptamers allows mod-

elling of an explicit sequence-fitness landscape. *Nucleic Acids Research*, 37(1):e6, 2009.

[150] W. Rowe, M. Platt, D. C. Wedge, P. J. Day, and D. B. Kell. Analysis of a complete DNA-protein affinity landscape. *Journal of Royal Society Interface*, 7(44):397–408, 2010.

[151] D. J. Jenkins and D. J. Stekel. A new model for investigating the evolution of transcription control networks. *Artificial Life*, 15(3):259–291, 2009.

[152] S. Kauffman and E. Weinberger. The NK model of rugged fitness landscapes and its application to the maturation of the immune response. *Journal of Theoretical Biology*, 141(2):211–245, 1989.

[153] P. Andolfatto. Adaptive evolution of non-coding DNA in drosophila. *Nature*, 437:1149–1152, 2005.

[154] C. P. Bird, B. E. Stranger, and E. T. Dermitzakis. Functional variation and evolution of non-coding DNA. *Current Opinion in Genetics and Development*, 16:559–564, 2006.